DOCUMENT RESUME

ED 307 980 PS 018 030

AUTHOR	Langhorst, Beth Hoover
TITLE	Assessment in Early Childhood Education: A Consumer's Guide.
INSTITUTION	Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Apr 89
CONTRACT	400-86-0006
NOTE	162p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Check Lists; Codes of Ethics; *Criteria; *Early Childhood Education; Guidelines; *Readiness; *Screening Tests; Test Norms; Test Reliability; Test Reviews; *Test Selection; *Test Use; Test Validity

ABSTRACT

Intended for use by early childhood practitioners who select and use early childhood assessment instruments, this guide provides information needed to judge instrument appropriateness and technical quality. The guide focuses mainly on standardized, broadly available instruments. Criteria for selecting instruments include publication after 1979, wide use, and provision of technical information. Other criteria concern whether the instrument targets children of 4-8 years of age and requires limited professional training for administration. The guide offers: (1) an overview of issues in early childhood testing; (2) a discussion of criteria for selection of instruments in general, and specifically, of those appropriate for developmental screening, readiness assessment, or instructional planning; and (3) discussions of the state of the art and prospects for the future, reviews of assessment instruments, and ways to choose an Early Childhood Education test. Appendices provide a code of fair testing, a list of reference works for early childhood assessment, and reviews or descriptions of over 50 early childhood assessment instruments. Reviews briefly describe the purpose of the instrument, test contents, administration format and procedures, scoring, norms, validity, reliability, utility, and availability. About 45 references are cited. (RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307980

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A CONSUMER'S GUIDE

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to:

In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view.

Assessment in Early Childhood Education

by

Beth Hoover Langhorst, Ph.D.

April 1989

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Northwest Regional Educational Lab.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PS 018030

Sponsored by OERI Office of Educational Research and Improvement
U.S. Department of Education

Northwest Regional Educational Laboratory
101 S.W. Main Street, Suite 500
Portland, Oregon 97204

Test Center

BEST COPY AVAILABLE

Assessment in Early Childhood Education
A Consumer's Guide

By

Beth Hoover Langhorst, Ph.D.

Northwest Regional Educational Laboratory
Test Center
Evaluation and Assessment
101 S W Main, Suite 500
Portland, OR 97204
(503) 275-9500

April 1989

Acknowledgments

The following persons reviewed the initial draft of the consumer's guide or contributed technical information. Their contributions were greatly appreciated.

Dr. Victoria R. Fu
Department of Family & Child Development
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Dr. Donald Ross Green
CTB/McGraw-Hill
Monterey, California

Dr. Randy Hitz
Early Childhood Education
Oregon Department of Education
Salem, Oregon

Dr. Linda McGarvey
CTB/McGraw-Hill
Monterey, California

Dr. Samuel J. Meisels
Center for Human Growth and Development
University of Michigan
Ann Arbor, Michigan

Dr. Wendy C. Roedell
Educational Service District #121
Seattle, Washington

Dr. Wendy M. Yen
CTB/McGraw-Hill
Monterey, California

April 1989

This guide is published by the Test Center of the Northwest Regional Educational Laboratory, a private nonprofit corporation. The work contained herein has been developed under Contract 400-86-0006 for the Office of Educational Research and Improvement (OERI), U.S. Education Department. The opinions expressed in this guide do not necessarily reflect the views of OERI and no official endorsement by the office should be inferred.

TABLE OF CONTENTS

1. Purpose for the Consumer's Guide 1
2. Issues in Early Childhood Testing 2
3. Selection Criteria for Early Childhood Assessment Instruments
   - How Should a Test's Technical Qualities be Evaluated? 5
   - Validity 5
   - Reliability 7
   - Norms 8
   - Adequacy of Information Available in the Manual 10
4. Selection Criteria for Screening Measures 11
5. Selection Criteria for Readiness Assessment Measures 15
6. State-of-the-Art and Prospects for the Future 18
7. Content of the Reviews of Assessment Instruments 20
8. How to Choose an ECE Test 22

References 23

Appendix A: Selection Checklist for Screening Instruments

Appendix B: Reviews of Screening Instruments

Appendix C: Summary Table of Screening Instruments

Appendix D: Selection Checklist for Readiness Mastery Instruments

Appendix E: Reviews of Readiness Mastery Instruments

Appendix F: Summary Table of Readiness Mastery Instruments

Appendix G: Reviews of Other Early Childhood Instruments

Appendix H: Summary Table of Readiness Mastery Instruments

Appendix I: Code of Fair Testing

Appendix J: Reference Works for Early Childhood Assessment

Index of Instruments by Category

List of Publishers Addresses

1. Purpose for the Consumer's Guide

Currently there is widespread concern over the increased use of standardized tests with young children. Two primary causes for this concern are the misuse of test results in making significant decisions that affect children's lives and with the lack of technically adequate assessment instruments available for legitimate testing uses (Meisels, 1987). The magnitude of the concern is reflected in public statements by the National Association for the Education of Young Children, the National Association of early childhood Specialists in State Departments of Education, the National Association of State Boards of Education, the National Black Child Development Institute, and the California State Department of Education School Readiness Task Force.

The National Association for the Education of Young Children "Position Statement on Standardized Testing of Young Children 3 through 8 Years of Age" puts the "burden of proof" on the test user.

> It is the professional responsibility of administrators and teachers to critically evaluate, carefully select, and use standardized tests only for the purposes for which they were intended and for which data exists demonstrating the test's validity. (p. 44)

To help lessen that burden, this guide is intended for use by early childhood educators or administrators who make decisions about the selection and use of early childhood assessment instruments. It is intended to provide the information necessary for users to judge the appropriateness and technical quality of instruments for early childhood assessment. Included are: (1) an overview of issues in Early Childhood Education (ECE) assessment, (2) a discussion of the criteria for selection of instruments appropriate for developmental screening or for "readiness" assessment and instructional planning, and (3) reviews or descriptions of more than 50 early childhood assessment instruments. The reviews briefly describe the intended purpose of the instrument, test contents, administration format and procedures, scoring, norms, validity, reliability, utility and availability.

This guide is not intended as a handbook on how to conduct early childhood assessments, nor to advocate a particular approach to assessment (e.g., standardized versus teacher-developed instruments). However, the primary focus is on standardized, broadly available instruments rather than observational, checklist or other less "formal" methods of assessment. It is intended to help clarify some of the sources of controversy and confusion surrounding the use of standardized assessment instruments, and to provide specific information about the appropriate uses and limitations of a variety of currently available measures.

Because there are hundreds of tests currently available that are designed for some type of early childhood assessment, it was necessary to limit the scope to assessment instruments that meet the following criteria:

- published in 1980 or later
- used widely (some earlier than 1980)
- provide technical information
- include the kindergarten age range (preferably ages 4 - 8)
- require limited professional training for appropriate administration

2. Issues in Early Childhood Testing

Current Controversies Over Assessment of Young Children

When standardized tests are used for purposes they neither were designed for, nor are technically adequate for, they are misused. This misuse has become a spotlight issue in Early Childhood Education, even to the point of a "call for a moratorium on the use of achievement tests in grades K-2" (Kamii, cited in Fromberg, 1989). While a large part of the concern stems from escalating academic expectations in kindergarten (Shepard & Smith, 1988), the real controversy has more to do with appropriateness of curriculum and instructional methods than with assessment alone. However, the filtering down of traditional elementary instructional practices into kindergarten has also brought increases in the inappropriate use of standardized testing.

The escalation of academic expectations in kindergarten, and increasingly in preschool, results from a downward shift of what were next-grade expectations, "trickling down" from competency standards or "promotional gates" in upper elementary grades. There are societal changes in child-rearing practices which also contribute. Much of what was traditionally the kindergarten curriculum is now taught before school by "Sesame Street," in preschools or by middle-class parents who, in turn, increase the accountability pressures on kindergarten teachers for evidence of children's academic progress (Shepard & Smith, 1988).

The stress on academic readiness and accountability in Early Childhood Education has given rise to what Fromberg (1989) has termed "a new set of 'three Rs': readiness testing, 'red-shirting' [delayed entry], and retention." Large numbers of school districts have begun to institute kindergarten admission and retention guidelines, as well as to provide extra-year "developmental" (pre-kindergarten) or "transitional" (pre-first grade) programs (Meisels, in press). Many school districts mandate a program of readiness testing before kindergarten or first grade entry. Some of the strongest objections raised by ECE practitioners and other experts to standardized "readiness" tests have to do with their use as the primary or sole criterion to determine kindergarten entry, placement in extra-year programs, and retention in kindergarten or first grade. The most important reasons that this is not a valid use of a readiness test have to do with the nature of "readiness" and have implications for equity issues.

The use of a readiness test to determine kindergarten entry or special program placement implies that it is possible to measure a level of readiness skill, and that a child below this level will not benefit from the instructional requirements of the kindergarten classroom. This conceptualization of the nature of readiness is based on the outmoded, unvalidated, but persistent underlying assumption that "readiness" is a function of maturation, and that the skills which traditionally make up readiness tests are an index of "developmental age." It is now widely acknowledged that if a child does not know letters, colors and shapes at the age of five, it is not more time, but more experience that is needed. Children at the age of three can know shapes, colors and letters and have a variety of emergent literacy skills if they have had exposure to these concepts. In terms of academic skills, "readiness" is achievement obtained before "formal" schooling. In practice, then, making decisions for kindergarten entry, placement, or retention based on academic readiness skills becomes a significant equity issue. A disproportionate number of children who are labeled as "unready" come from low-income and culturally varied groups (Fromberg, 1989; Hilliard, 1985; Abidin, Golladay & Howerton, 1971).

The validity of using readiness tests for decisions on kindergarten entry or placement has, as Shepard and Smith (1986) suggest, become entwined with the validity of the decisions in which they are involved, and with effectiveness of special programs. While there is correlational evidence that some readiness tests do predict success in kindergarten or later achievement, it does not follow that children who perform poorly on readiness tests will benefit by being kept out of kindergarten. Those children may be the ones who will benefit the most if provided a flexible, appropriate kindergarten curriculum (NAECS/SDE, 1987).

Should Young Children Be Assessed?

Teale (1988) suggests that educators now stand at an important crossroads on the issue of assessment in early childhood education. Because issues of assessment are intertwined with the still evolving issues of curriculum escalation and developmental appropriateness, there are no quick or easy resolutions with kindergarten admission, retention and extra-year policies. However, there are legitimate and important reasons for assessment in early childhood. What legitimizes the assessment of children is that the results are used for their benefit.

Two major reasons for large-scale assessment of young children are:

- screening to identify children at risk for potential learning problems and in need of further, more intensive evaluation
- assessment of readiness for a specific academic program, to facilitate instructional planning and curriculum, both on an individual and school policy level

Along with the increase in readiness testing there has been an increase in large scale screening programs for young children. The use of standardized "readiness" tests in "screening" for school entry or placement has fed the confusion regarding the distinction between screening for potential learning problems and assessing "readiness" for a particular instructional program. While readiness assessment should be concerned with the skills a child has acquired (e.g., letter names), screening instruments focus on the underlying abilities to acquire those skills (e.g., visual and auditory discrimination).

> Developmental screening is a brief assessment procedure designed to identify children who, because of the risk of a possible learning problem or handicapping condition, should proceed to a more intensive level of diagnostic assessment. Screening serves as the first step in an evaluation and intervention process that is intended to help children achieve their maximum potential. (Meisels, 1985, pg. 1)

Early childhood educators, policy makers, and legislators agree that, in solving learning problems, early identification and intervention are important. Developmental screening for 3- to 6-year-olds is already mandated by 25 states (Meisels, 1986). Developmental screening will continue to increase as a consequence of Public Law 99-457, which amended the Education for All Handicapped Children Act (PL 94-142) to include handicapped preschool children.

Because of the controversy previously described, there is much less consensus on the issue of readiness assessment. However, the same argument that makes readiness tests invalid as selection devices strongly supports their use in individualized curriculum planning. The position of the National Assocation for the Education of Young Children is that

> ... [A]ssessment of individual children's development and learning is essential for planning and implementing developmentally appropriate programs, but should be used with caution to prevent discrimination against individuals and to ensure accuracy. (Bredekamp, 1987, pp. 12-13)

Beyond the benefits for individual children, the results of screening and readiness assessment can be important for decisions at school, district, or state levels. These include policy decisions, and planning for funding staff development and curriculum development. Even when assessment instruments are used appropriately, however, the value of the information is dependent on the technical quality of the measure.

> Accurate testing can be achieved only with reliable, valid instruments, and such instruments developed for use with young children are extremely rare. (Bredekamp, 1987, p. 13)

Clearly there is a need for greater care in the selection process and stronger adherence to high standards of technical quality in early childhood assessment as a whole. Meisels (1987) notes that there is a nationwide proliferation of screening tests, many of which are locally developed, that have never been assessed in terms of reliability, validity, or other criteria for evaluating screening tests. He cites surveys of school districts in New York State (Joiner, 1977) that found only 16 out of 151 screening tests or procedures were even marginally appropriate, and another in which fewer than 10 out of 111 tests being used for preschool, kindergarten and pre-first grade programs were appropriate in terms of the age group and purpose (Michigan Department of Education, 1984).

Persons without a background in assessment may be tempted to assume that if an assessment instrument is published and/or widely used, a careful scrutiny of technical details is not needed. Unfortunately, such evidence of "cash validity" (Salvia & Ysseldyke, 1988) is often misleading. The brief summary of technical terms and issues in the next section is intended to make the technical review process as painless as possible.

The basic types of assessment instruments to be considered in this guide reflect the major uses of large-scale assessment described above, screening to identify children at risk of potential learning problems, and mastery of readiness concepts used for instructional planning. The selection criteria specific to screening instruments are provided in a checklist format in Appendix A. A similar checklist is provided for readiness mastery instruments in Appendix D. The technical standards apply as well to measures that assess cognitive ability, developmental milestones, or motor skills.

3. Selection Criteria for Early Childhood Assessment Instruments

How Should a Test's Technical Qualities Be Evaluated?

The American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME) have jointly published the *Standards for Educational and Psychological Testing* (1985) as a guide for both test producers and test users. The NAEYC position statement on standardized testing acknowledges and endorses the APA standards.

> Standardized tests used in early childhood programs should comply with the joint committee's [APA, AERA, NCME] technical standards for test construction and evaluation, professional standards for use, and standards for administrative procedures. This means that no standardized test should be used for screening, diagnosis, or assessment unless the test has published statistically *acceptable* reliability and validity data. (NAEYC, 1988, p. 43)

What constitute "acceptable" reliability and validity data are somewhat different for screening than for readiness assessment purposes. Selected issues which are particularly relevant for screening and for readiness assessment are outlined below and in the separate selection checklists. Parts of the APA guidelines relevant to issues of test use in education have been summarized for a general audience in the *Code of Fair Testing Practices in Education* (1988; Appendix I). A more detailed discussion is presented in the *Standards for Educational and Psychological Testing* themselves. References which might be more understandable for the "lay" reader include the *Handbook for Measurement and Evaluation in Early Childhood Education* (Goodwin & Driscoll, 1980) or *Assessment in Special and Remedial Education, Fourth Edition* (Salvia & Ysseldyke, 1988). A list of reference works on assessment in early childhood education is provided in Appendix J.

Validity

In general, validity refers to the extent to which the test fulfills the purpose for which it is intended. For example, the most important criteria for a screening measure is that it accurately distinguishes the children in need of further assessment from those who are not. Validity is the most important attribute of a test. Other aspects such as reliability, adequacy of norms, or lack of bias are all necessary but not sufficient conditions for validity (Salvia & Ysseldyke, 1988).

Validity always pertains to a specific use of a test. A test may be valid for readiness but not for screening. It is the responsibility of the test developer to present evidence for the types of validity and reliability most appropriate to the use of the test for the purposes it is intended. There are three major types of validity: content, criterion-related and construct.

Content validity refers to the extent to which test items represent the larger body of content or "domain" the test is intended to measure. In judging the content validity of a test the user should consider three things:

- the appropriateness of the items in terms of what is being measured
- the completeness of the item sample (covers all important areas of content)
- the way in which items assess the content (Salvia & Ysseldyke, 1988)

An assessment instrument for early childhood should reflect a theoretically accurate picture of the skills and knowledge that students are developing during this period (i.e., appropriate and complete content; Teale, 1988). In the case of "reading readiness," for example, new standards for content validity are evolving as the research on emergent literacy broadens our understanding of the development and the integration of the component skills involved in learning to read. Matching the content of a readiness test to the curriculum content of a specific instructional program is an important step in assessing content validity.

The way in which content is assessed should also be sensitive to developmental and personal characteristics of young children that relate to testing, such as attentiveness (Teale, 1988). The format of a test in terms of what response is required of the child (e.g., paper/pencil, pointing, verbal), how long it takes to administer, or how instructions are worded are all factors which affect whether the items really measure what is intended and are important validity considerations. The content and format should be consistent with the way in which children interact with, think and learn about their environment, i.e., "developmentally appropriate."

Typically, content validity is addressed by test developers through a logical analysis of test content by subject-matter experts. Sometimes the experts are the test developers themselves; sometimes test contents are reviewed by teachers or research experts. A test manual should give a clear definition of the universe or domain of content represented by the test. It should describe procedures which were followed to select items representing that domain, explain why those procedures were valid, and present the qualifications of any subject-matter experts who made judgements about the content.

Criterion-related validity refers to the extent to which a child's performance on an assessment measure can be used to estimate performance on a criterion measure, whether it is future performance (**predictive validity**) or performance at the same point in time (**concurrent validity**). Criterion-related validity is typically judged by the strength of the correlation coefficent between the assessment measure and the criterion measure, or by how well a classification based on the assessment measure (e.g., at risk/not at risk) matches the actual outcome based on the criterion measure.

Predictive validity is often more valuable information than concurrent validity in instructional planning or screening. It is much more costly to conduct a study of predictive validity, however, so many authors settle for evidence of concurrent validity. For either type of criterion-related validity, the confidence in the evidence of validity is limited by the confidence in the validity of the criterion measure, the number of children in the study, and whether the children in the study sample are comparable to the children who will be assessed.

A good test developer will present evidence for the instrument's ability to predict outcomes for students from a variety or nationally representative sample of backgrounds. A measure that predicts the future school performance of a group of upper-middle class Kindergartners in Vermont may not do as well predicting the performance of Hispanic children in inner-city New York. If predictions from a screening test are different according to, for example, preschool experience, and if the same standards are used without regard to preschool experience, then the test results will be biased for some children. Unfortunately very few early childhood tests deal with this issue, partly because such studies are very costly.

Construct validity refers to the extent to which the test measures a theoretical construct or trait. Examples of constructs are intelligence, endurance, creativity or self-esteem. The test developer should present the conceptual framework, however simple, that clearly specifies what the test intends to measure, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables. For example, if a test is based on a construct that theoretically changes with age, evidence should be presented that performance on the test does, in fact, differ for different age groups.

Factors that can affect validity include reliability, administration errors, whether or not the test is administered in the primary language, or norms that are not representative.

Reliability

Reliability refers to the degree to which scores are free from error of measurement, i.e., consistent, dependable and repeatable. It usually takes the form of a measure of the consistency of test scores over time (**test-retest**), or over different test-givers (**inter-rater**). **Test-retest** reliability, giving the same test twice with a brief intervening interval, establishes confidence in the stability of assessment results. When a test provides cutscores for making important decisions such as referral for diagnostic evaluation, the reliability of the cutscores should be addressed in addition to the reliability of total or subtest scores. The correlation between test scores used to establish reliability is called the reliability coefficient.

If a test relies on an observer rating children's behavior (rather than direct interaction with the child), the criteria for scoring each item in the test manual should be clear and explicit so that different observers will score the same behaviors in the same way. The manual should provide evidence of **inter-rater** reliability, i.e., statistical evidence that two observers rating the same child come up with very similar ratings.

Internal consistency is another measure of reliability often measured by the correlation of scores from one half of the test items with scores from the other half (split-half) or by examining the correlation between a score on one item and the total for the rest of the items. Consistency among items within a test or subtest is evidence that the test measures a single construct; however, it is not evidence of construct validity because it does not provide information on what single construct is being measured. If the test has more than one equivalent form, it is important to establish the consistency of scores between the forms (**alternate forms** reliability).

It is important to consider the number and the age range of children who participate in reliability studies. Particularly when young children are involved, the stability of test scores may vary with the age group. Reliability coefficients should be provided for each separate age or grade group for which there are test scores. (For example, in six-month intervals if scores are presented in six-month intervals.) Sometimes, when a study sample is small and the age range is large, reliability coefficients will be presented only for the group as a whole. This increases the amount of variation in scores due to age in relation to the variation due to errors of testing and artificially inflates the reliability coefficient. On the other hand, reliability coefficients may sometimes be reduced by the lack of variability in test scores, due to a "ceiling" effect.

Norms

For a norm-referenced test, the meaning of a child's performance is based on a comparison to the performance of other children. Raw scores (the total number of items answered correctly) can be hard to interpret without reference to some standard of performance. Norm-referenced tests measure mastery of specific skill relative to how the children in a reference group performed, not in relation to a absolute level of mastery of those skills. Norm-referenced tests can be contrasted with criterion-referenced tests in which performance is evaluated relative to a predetermined level of mastery of specific skills (the "criterion"). Some norm-referenced tests also provide criterion-referenced information.

One of the most important criteria in selecting a norm-referenced test is that the reference group (often called the "standardization sample") is representative of the population being assessed. Even for a criterion-referenced test, it is important to examine the representativeness of the population of children who participated in piloting, reliability and validity studies. Many times the normative or reference group is national, but it also may be local or statewide, depending on the intended use of the scores.

It is important that there are at least 100 children in each age or grade interval on which scores are based, to ensure the stability of the normed scores (Salvia & Ysseldyke, 1988; Sattler, 1988). The overall representativeness of the sample should be reflected in each of the score groups. The sample may be balanced geographically but if all the three-year-olds were from one state and all the four-year-olds from another, the sample would not be representative in any practical sense. The minimum of 100 children per age or grade interval used for scoring is frequently violated in the standardization of early childhood instruments because of the time involved in individual administration.

Standard scores, developmental ages, grade equivalents and percentiles are among the most frequently used normative standards of comparison. These are "derived" from the raw scores as developmental scores or scores of relative standing. Outlined below are a number of issues, regarding the way in which scores are developed and used, that have important consequences for the interpretation and use of test results.

The most frequent **developmental score** used for early childhood assessment is the age equivalent or "developmental age." An age equivalent score represents the average or middle raw score of children at a particular age level, usually expressed in years and months (e.g., 4-0). Test developers often group children by six-month intervals in order to establish age-related scores for early childhood tests. The "developmental quotient" is sometimes derived from the developmental age (the developmental age, divided by the chronological age, multiplied by 100) in order to quantify the rate of development.

The use of developmental-age scores persists despite widespread criticism that they are easily and often misinterpreted by professional as well as lay persons (Cronbach, 1970; Goodwin & Driscoll, 1980; Allen & Yen, 1979; Burkett, 1986). In support of the opinion that "*developmental scores should never be used*" (Salvia & Ysseldyke, 1988), professional organizations such as the International Reading Association, the American Psychological Association, the National Council on Measurement in Education and the Council for Exceptional Children have "very negative official opinions" about developmental scores and quotients.

The problem with developmental age scores which is most critical for the purposes of early childhood assessment is that they imply that there is such a thing as an "average" five-year-old. This creates false standards of performance. The 5-0 "developmental age" is based on a range of scores as would be expected from the wide variation and rapid growth in development in early childhood. Fifty percent of the 5-0 age group in the norming sample performed at or **below** that score.

Developmental-age scores encourage comparisons with inappropriate age groups (Sattler, 1988). In practical terms, a bright three-year-old who passed enough items to get a developmental-age score of 5-0 does not have the same range of skills and experience, and is not ready for the same instructional climate as a five-year-old with a score of 5-0. Developmental-age scores incorrectly imply that what is being measured varies in equal units between age groups. Because the rate of development is not constant over the early years it would not be expected that every skill will develop to the same degree in the six month difference between 3-0 and 3-6 as in the six months between 5-0 and 5-6.

Raw scores are frequently transformed into **standard scores** which, while widely used, are also open to misinterpretation. One widespread assumption is that standardizing scores makes them readily comparable with standardized scores from other tests. However, the scores should legitimately be compared only to standardized scores of tests with a similar distribution of raw scores. Many measures promote the comparison of standardized scores between subtests of one measure, or between different measures, without adequately dealing with issues of comparability.

Among **scores of relative standing**, the most easily and accurately interpretable standards of comparison are percentiles and percentile ranks. Percentiles are calculated by dividing the scores in the normative group into 100 equal groups. Percentile ranks indicate the percentage of children scoring *at or below* a given score. For example, if a child's raw score corresponds with a percentile rank of 55, 55 percent of the normative sample scored at or below that raw score.

One often-misunderstood aspect of a norm is that it "is not a standard or a goal to be reached" (Goodwin & Driscoll, 1980). Only in Lake Woebegon are all the children above average. Elsewhere, by definition, half the children are at or below the 50th percentile. In any score of relative standing it is extremely important to compare children to an appropriate reference group, in terms of their home and preschool experiences.

Teachers often object to norm-referenced tests because many young children will "fail" many items. This is necessary in a test that is designed to measure well across a wide range of ability. If most children passed most items, their true level of skills could not be determined. Such a "ceiling effect" limits the usefulness of the information to be gained from assessment. It is important to protect children from situations where they fail because of inappropriate levels of difficulty or inappropriate test formats. However children cannot get a sense of accomplishment if it is impossible to fail. It is not beneficial to "protect" children from knowledge of their own limitations, nor to protect administrators and policy makers from knowing what and where the real needs are.

One hallmark of a "standardized" test is that the format for administration, including instructions, materials and setting are clearly specified so that they can be consistent for each child. The appropriateness of the norms for a given child depend on how well the assessment circumstances matched the standardization circumstances. The test manual should present directions for administration and scoring that are explicit and easy to duplicate.

Adequacy of Information Available in the Manual

In addition to clear directions on how to administer and score a test, the manual should be the best source of the information needed to evaluate validity, reliability, the representativeness of the norms, and to determine for what purposes the test can legitimately be used.

The following sections discuss aspects of these technical issues that are particularly important for screening or readiness assessment instruments. Separate checklists are provided that cover the specific selection criteria for screening (Appendix A) and for readiness mastery (Appendix D) instruments. These criteria were used in the reviews of measures for this guide and should be useful in reviewing other related measures.

4. Specific Selection Criteria for Screening Instruments

Validity

Content validity: appropriateness and completeness. The purpose of developmental screening is to identify children who may have a learning problem or handicapping condition that could affect their potential for learning. A developmental screening test should be brief, inexpensive, norm-referenced with clear, standardized administration and objective scoring procedures, and broadly focused over a range of areas of development, including speech, language, cognition, perception, affect, gross and fine motor skills, and personal/social behaviors. (Meisels, 1985, 1988).

Screening procedures should also include parental input, vision, hearing and health assessment (NAECS/SDE, 1987). Specific vision, hearing and health assessment measures are not within the scope of this Consumer's Guide but some of the reference works listed in Appendix J discuss such instruments. The selection checklist for screening instruments in Appendix A provides an outline of content coverage which has been used to evaluate the content of the measures reviewed in Appendix B. Information about how specific content was chosen and evaluated in test development, particularly evidence of the predictive validity of specific areas of content, should also be examined.

Many school systems, failing to find one instrument that covers all of the areas they want to screen, put together their own screening system, often with bits and pieces of many different screening tools. Information provided in test manuals or research studies about the validity or reliability of a particular assessment measure, unless it is presented for specific subtests, cannot be applied to just part of the test used in isolation. When a test is developed from bits and pieces of others, evidence for validity and reliability must be established for the "new" test.

Criterion-related validity. The assessment of how accurately a screening test classifies children in terms of risk (and therefore need for further assessment) is one of the most significant criteria for selection, because of the importance of this information for individual children. The purpose of a screening device is to provide information for a yes/no decision on whether or not a child will be referred for diagnostic assessment. A cutoff point must be established, therefore, that differentiates potentially "at-risk" individuals from those not in need of further evaluation. The criterion-related validity is measured in terms of the accuracy of that decision.

The **sensitivity** (accuracy in identifying all "at-risk" children) and the **specificity** (accuracy in identifying all children *not* "at risk") of the test can only be determined by comparing screening results of children above and below that cutoff point to the results of some outcome measure such as the results of a diagnostic test. Figure 1 illustrates the procedure for determining **specificity** and **sensitivity** as well as false positives (**overreferrals**) and false negatives (**underreferrals**).

Figure 1. Criterion-Related Validity for Screening Measures

		Follow-up assessment **OUTCOME**	
		Intervention needs **+**	No Intervention needs **-**
Screening Test	At risk: Refer for evaluation **+**	True Positives **a**	False Positives *overreferrals* **b**
	Not at risk: Do not refer for evaluation **-**	False Negatives *underreferrals* **c**	True Negatives **d**

Sensitivity: The proportion of children at risk who are correctly identified.

$$\frac{a}{a + c}$$

Specificity: The proportion of children not at risk who are correctly excluded from further assessment.

$$\frac{d}{b + d}$$

Adapted from Meisels (1985). *Developmental Screening in Early Childhood: A guide.* Washington, DC: NAEYC.

The degree to which **specificity** (accurate identification of true negatives; i.e., correctly unreferred) and **sensitivity** (true positives, i.e., correctly referred) are important depends on the negative consequences of an error in screening. The consequences for the child identified as having a potential problem, when in fact diagnostic tests and future performance do not indicate a problem (false positive), are considered less serious than those related to not identifying, and therefore not evaluating or intervening with, a child who actually does have a learning problem (false negative). The consequences of not identifying a child who needs intervention are significant for the school system as well as the child. A learning problem that is not identified early may become exacerbated and more difficult and costly to remediate. Also, a high rate of underreferrals means that future needs in terms of planning for special programs and staffing will be underestimated.

Because there are significant consequences for misidentifying children, screening tests should only be used if they have a high degree of sensitivity and specificity, along with evidence of reliability. Meisels (1988) suggests a minimum criteria of 80 percent for sensitivity and specificity. He also points out, however, that because of the necessary brevity and broad scope of screening instruments, they cannot be expected to be 100% accurate.

Correlational studies showing a strong relationship between screening and outcome do not give information about accuracy for individuals and should not be accepted as replacement for sensitivity and specificity as evidence of validity.

> ... [T]est producers are strongly encouraged to present data concerning the proportion of at-risk children correctly identified (test sensitivity) and the proportion of those not at-risk who are correctly found to be without major problems (test specificity). (NAEYC, 1988, p. 43)

Reliability

Evidence of both **inter-rater** reliability and **stability** is important for screening tests. Evidence of internal consistency is much less important and is not a substitute for other forms of reliability. Evidence of the reliability of classification based on cutscores should be provided.

Norms

Screening tests are essentially norm-referenced tests because the decision point for referral, the "cut score," is determined in relation to a normative group. The normative group may be national or local, but it should be as similar as possible to the population being screened. If the normative group is very dissimilar from the screening population, the norms and any cut score based on those norms are not applicable. Evidence for sensitivity and specificity is only true for the specific cut score used in the validity study. The number and characteristics of the children in the study sample also influence the degree of confidence in validity evidence.

Some screening devices have more than an either/or cut score, and different levels of "risk" status can be used to prioritize the need for referral. Schools or districts can determine a cut score, or series of priority scores, that are specific to their populations. Many school districts cannot realistically evaluate in a timely manner all children referred from a screening program. If studies of the validity of classification results are conducted, locally developed cut scores or priority scores can be adjusted to yield the maximum sensitivity and specificity that are logistically feasible. It should also be kept in mind that cut scores may change as normative groups change.

The directions on how to administer and score the test should be presented clearly, so that the screening conditions can be as similar as possible to the conditions under which the test was normed. The potential user should determine how dependably these conditions can be replicated in aspects such as the facilities needed, the time for administration, and the training needed by administrators.

Limitations of Screening Instruments

Even a well developed screening instrument can be used inappropriately. Meisels (1985) lists the following limitations of screening instruments that should be kept in mind when selecting a screening instrument and designing a screening program.

- The data from screening instruments should not be used as diagnostic/assessment information.
- A screening instrument is not identical to a school entrance test.
- Screening tests are not IQ tests (i.e., they are not measures of a child's *overall* cognitive functioning).
- Screening results should not be used to label a child. No screening test is comprehensive enough to identify a child as having special needs.
- Screening tests should not be used in multicultural/multilanguage communities if they are not sensitive to cultural differences or the effects of bilingualism.
- Screening should never be performed in isolation. It should always be performed within the context of a program of assessment, evalution, and intervention.

5. Specific Selection Criteria for Readiness Mastery Instruments

Validity

Content validity: appropriateness of content and method. The purpose for readiness testing is to assess the level of existing knowledge and skills in order to individualize curriculum planning and/or to provide information on groups of students for school policy decisions such as curriculum restructuring. The appropriateness of the content of a specific assessment measure should be judged in terms of how useful the information it provides will be in terms of the specific decisions that need to be made.

In the current period of re-examination of curriculum content and instructional methodology in Early Childhood Education, the definitions of constructs such as readiness are changing. This makes it more even more crucial to examine how well the content of a given instrument reflects current theory and practice. The focus on developmental appropriateness leads to a closer examination of the manner in which specific content is being measured. For example, is the response required of the child age appropriate? Does the item address an isolated skill or put the information in a context with which the child is familiar? While screening measures should cover a broad range of content areas, assessment for instructional planning can involve content coverage as broad as the issues involved in restructuring the early elementary curriculum, or as narrow as an individual teacher's assessment of counting skills.

Very often teachers' judgments are used as the criteria for evaluating the validity of a measure. This leads one to ask whether formal assessment instruments are useful to teachers. Teachers commonly create informal asssessment tools and procedures for their day-to-day instructional planning. The type of information provided by a formal readiness test can be useful for instructional planning for a number of reasons. One of the most significant reasons is that it provides a tool to structure the teachers observations so that the same information is gathered consistently for all children.

In an ideal situation, a teacher would be well trained in current theory about young children's growth and development, have a limited number of children in the classroom, and know them all well. However, not all teachers have equal experience and training, nor do they have the same amount of interaction with all children in a classroom. Yet the teacher is expected to be sensitive to the specific needs of individual children and to be able to communicate about those specific needs with parents, next-grade teachers, and other professionals. A formal measurement gives objective, consistent criteria to evaluate performance not only for planning instruction or further evaluation, but also to substantiate the teacher's judgments when communicating with parents and other professionals. The use of a test for these purposes does not replace or lessen the value of the teacher's judgement.

The selection checklist for readiness mastery instruments (Appendix D) provides an outline of content coverage which has been used to evaluate the content of the measures reviewed in Appendix E. The content categories were derived from the World Book, Inc., survey of more than 3,000 kindergarten teachers throughout the United States and Canada on the skills and knowledge a child needs to begin kindergarten successfully. They are provided here as a quick comparison of the scope of test content. Information about how specific content was chosen and evaluated in test development, particularly evidence of the predictive validity of specific areas of content, should also be examined.

A perfect match of test content to curriculum is not necessary or to be expected. The closeness of the match should be weighed against other aspects of quality, such as evidence of reliability and validity. There is similarity among kindergarten curricula and teacher judgements (as evidenced by the World Book survey) which is reflected in the similarity of content among readiness tests. Nationally normed readiness tests are typically based on surveys of preschool and kindergarten curricula, and are generally more technically adequate than locally produced measures. They have the added benefit of allowing teachers and administrators to compare the relative skill levels of their children to those across the nation. Such information can be useful in justifying and evaluating early intervention programs.

Criterion-related validity. One way to establish that a test is measuring what it intends to measure is to compare performance on the test with performance on a criterion test for which validity has already been established. For example, in order to establish the validity of the Lollipop Test as a measure of readiness, the test's author compared the performance of the same group of children on the Lollipop Test and on the Metropolitan Readiness Test (MRT) (Chew & Morris, 1984). This type of evidence of concurrent validity is most useful when performance on one test is intended to be used to estimate performance on the criterion test (Allen & Yen, 1979). In the case of the Lollipop Test the author wanted to establish that his shorter test measured readiness as well as the longer MRT.

Sensitivity and specificity are not characteristics of most readiness tests, however, and that is one primary reason most readiness tests are not valid screening devices. The issue that weakens such evidence of predictive validity for readiness tests is that children who perform poorly may profit more from school programs than children with higher initial skills, because they have more to gain. This results in a high rate of "false positives" that is discriminatory toward disadvantaged children. Readiness tests are used most appropriately to "describe *child entry characteristics*; they are not intended to *predict child outcomes*" (Meisels, 1987).

Reliability

Evidence of *inter-rater* reliability is important for readiness assessment only if the response requires a rating of the child's behavior by teachers and/or parents. When the information is being used for program planning, evidence should be required that what is being measured is a stable characteristic of the child. However, evidence of *stability* over time (*test-retest* correlations) should be expected only for short time intervals (e.g., weeks); long-term stability would not be expected for readiness skills during a period of rapid development and information acquisition. Evidence of *internal consistency* provides confidence that items within a content area are measuring the same thing.

Norms

Readiness mastery measures are essentially criterion-referenced tests because they are designed to assess specific curriculum goals. They can also be norm-referenced. Because differences in mastery of readiness skills in kindergarten are due more to experience than to age, scores of relative standing would be more appropriate standards of comparison than developmental age.

As with screening tests, the normative group may be national or local, but it should be as similar as possible to the population being screened. Parental education is a crucial factor influencing differences in mastery of readiness skills, but is often neglected in the development of normative information.

Limitations of Instruments Measuring Mastery of Readiness Skills

No matter how well it meets standards of technical quality there are limitations on the information a single assessment instrument can provide.

- Particularly because of the rapid growth and development of young children, test results must be seen as a "snapshot" view--an indication of performance in a limited context, in a limited time frame.
- Assessment of mastery of early skills is not a measure of a child's *overall* cognitive functioning, nor an index of "developmental level."
- A child's performance will be influenced by motivation and temperament factors which are not addressed by readiness measures, but are an important part of the total picture of the child within an instructional context.
- Assessment results should not be used to label a child. Readiness measures provide information on a child's strengths and weaknesses in terms of specific knowledge, but not on strengths and weaknesses in ability to acquire that knowledge.
- Measurement of readiness skills is influenced by cultural differences and the effects of bilingualism. The results should be interpreted with caution in communities that are multicultural/multilanguage
- Readiness assessment should never be performed in isolation. It should always be performed within the context of instructional planning, whether in the classroom or on a school, district or state level.

6. State-of-the-Art and Prospects for the Future

State-of-the-Art

There are many reasons for the current level of concern about the use of standardized testing. These include inappropriate curriculum and instructional methods, confusion about the distinction between screening for potential learning problems and "readiness testing," and the use of instruments for purposes for which most were not designed nor are technically adequate. The misuse of early childhood assessment instruments by many consumers occurs when they do not clarify the goals of the assessment process sufficiently, do not take into account the limitations of assessment, and do not understand or use standards of technical quality as selection criteria.

This situation is exacerbated by the lack of adherence to standards of technical quality on the part of test developers. Perhaps the most serious issue, because of the widespread practice of screening and the significance of the consequences of errors in identification, is the lack of evidence for the validity of referral cutscores. Many screening measures propose cutscores but offer no rationale and no evidence of sensitivity or specificity. Evidence that the test scores were stable over time (test-retest reliability) is often based on very small samples with a limited age range. Evidence of the stability of classifications based on cutscores is very rare.

In addition, the normative groups on which percentiles, standards scores or cutscores are based are often poorly described, not representative of the general US population, and inadequate in terms of numbers of children in each age or grade interval. The most frequent factor known to influence children's skills but not considered in reference groups is the child's educational experiences at home (indexed by parental education) and at preschool. Only a few measures have separate norms by socioeconomic level, and none of those reviewed consider the amount of preschool experience as a separate factor.

While the consumer of early childhood instruments need to be aware of technical issues, it is the responsibility of the test publishers to provide sufficient technical information on which to judge the quality of the test and the appropriateness of the normative group. Test publishers clearly need to make a greater effort to provide detailed descriptions of how tests are developed and normed, as well as strong evidence of the reliability and validity of scores.

Educators now stand at an important crossroads, not only on the issue of assessment in Early Childhood Education (Teale, 1988), but on the still-evolving issues of academic curriculum escalation and developmental appropriateness, kindergarten admission, retention and extra-year policies. In the current period of re-examination of curriculum content and instructional methodology in Early Childhood Education, the definitions of constructs such as readiness are changing. The questions of *what* specific curriculum content should be assessed as well as *when* and *how* and with *whom* it can be appropriately assessed cannot be easily or quickly answered.

New definitions of "reading readiness" are evolving as the research on emergent literacy broadens our understanding of the development and the integration of the component skills involved in learning to read. Even though new measurement approaches often emerge along with new developments in research and practice, it takes time to establish instruments with proven validity and reliability.

The Future of Early Childhood Assessment

The current controversies regarding assessment practices have led to a greater recognition that "readiness" is more than isolated academic skills, and that curriculum and instructional practices should adjust to the various needs of children entering school rather than "screening out" children who do not meet the requirements of the curriculum. This does not rule out the benefit of using standardized tests for instructional planning and evaluation purposes. However, it should result in a closer examination of the manner in which specific content is being measured. For example, is the response required of the child age appropriate? Does the item address an isolated skill or put the information in a context with which the child is familiar?

It is important that the over-reliance on standardized testing and legitimate concerns over misuse and technical inadequacy do not lead to throwing the baby out with the bathwater. In response to the search for more "developmentally appropriate" assessment, there is an increased focus on such methods as observation and the collection of work samples. Spodek (1988) cautions that the exclusive focus on developmental appropriateness should not lead us to lose sight of fundamental questions about the academic content of curriculum. While there is movement toward innovative middle positions, such as structured performance assessments in the area of emergent literacy (Teale, 1988), there are still legitimate reasons for standardized assessment of young children. Particularly when such major changes as restructuring and integrating more academic curriculum and practice in early elementary grades are being called for, there is legitimate need for information on the level of mastery of specific early academic or "readiness" skills.

There will be a tremendous influx of new and revised instruments for early childhood education coming into the market. Some will address some of the concerns about developmental appropriateness in assessment. What early childhood assessment consumers must require of such products is a greater focus on the interface between teaching and assessment, a greater acknowledgement of the parent as a source and consumer of assessment information, and clear, consistent evidence of the quality of tests intended for assessment of young children in terms of content, format, the size and representativeness of normative groups, as well as strong and consistent evidence of validity and reliability.

7. Content of the Reviews of Assessment Instruments

The selection criteria checklists, reviews of early childhood assessment instruments and summary tables are in separate appendices. This was done because it may be useful to make copies of the different appendices for the use of test review committees, depending on the type of measure wanted and the stage in the review process.

The reviews have been done with the previously outlined issues and standards in mind. Checklists are provided in Appendices A and D outlining specific criteria for test selection in screening and mastery of readiness skills. The reviews describe the stated purpose for the instrument, the instrument format, content, administration, and available scores, and rate the technical quality in terms of evidence of validity, reliability, and adequacy of norms.

Evidence of validity and reliability have been rated according to those aspects most important for the test purpose. Evidence of stability of test scores (test-retest) and inter-rater reliability, for example, is valued more highly than is evidence of internal consistency. A rating of "poor" indicates that the instrument does not meet the most important criteria and its use would not be recommended for the stated purpose. A rating of "fair" indicates that the measure fails to meet some of the criteria completely, or meets them in a manner that limits applicability to specific populations, but does as well as most available instruments. A rating of "good" indicates that the instrument meets all the most important criteria, while those rated "excellent" provide not only all the important evidence, but also provide information that enhances the interpretation and utility of test results.

The ratings also take into account the number of children in the reliability and validity studies, whether evidence is presented for all age groups and all relevant scores (e.g., cutscores as well as total scores), and the demographic characteristics of the group. The general standard for such samples, as well as norm groups, is that they be representative of the US population in general or an explicitly defined special reference group. It is noted if validity or reliability coefficients could have been inflated by including a large age range in one correlation, or deflated by ceiling effects. If the evidence is strong but limited in applicability, the word "limited" may be used to qualify the rating. The ratings are also qualified if the evidence is good for one age group but not another.

Appendix B contains reviews of instruments designed to screen for potential learning problems, with a summary table in Appendix C. Appendix E contains reviews of instruments to measure the mastery of readiness skills, with a summary table in Appendix F. Appendix G contains reviews of other early childhood instruments that are widely used but do not fit into either the screening or readiness category (i.e., developmental inventories and instruments of cognitive maturation), with a summary table in Appendix H. Also included in the summary tables are instruments which are not reviewed in full. The absence of a full review is not meant to imply quality. The instruments which were reviewed were selected to represent a variety of what is available in terms of scope and quality. While some instruments were not reviewed because of poor quality and limited information, time constraints also played a role.

The following keys to ratings for norms, reliability and validity are provided with the summary tables.

NORMS:	*Ratings on norming studies (value judgement implied)*
None:	no normative information is given
Poor:	some information but limited applicability
Fair:	some standards of comparison (e.g., means of research sample)
Good:	norms based on good sized, representative sample, or lots of other relevant information regarding appropriate populations for use
Excellent:	norms based on a representative, national sample and relevant information about applying norms or norm-referenced scores.

RELIABILITY:	*Reliability ratings (value judgement implied)*
None:	no reliability information is provided
Poor:	all reliability coefficients (**r**) below .70 or an important type of reliability was not examined
Fair:	at least one reported **r** is greater than .70; or **r** was greater than .80 but evidence was limited in applicability
Good:	total **r** is greater than .80; most subtests have **r** greater than .75
Excellent:	several kinds of reliability reported; total **r** is greater than .90; most subtest scores greater than .80

VALIDITY:	*Validity ratings (value judgement implied)*
None:	no validity information is provided
Poor:	information is of very limited applicability
Fair:	most important aspects of were addressed but evidence was moderate or weak; or was strong but limited in applicability
Good:	consistent evidenct of validity, or strong but limited evidence of the type of validity most appropriate for the intended test use
Excellent:	strong evidence and a base of research on the instrument

8. How to choose an ECE test

No one instrument will meet all the criteria outlined in the checklists for screening or readiness assessment. The most important criteria for test selection is how useful the information will be in the specific planning or decision making process you are addressing. Will the information lead to beneficial changes in the educational development of children? Ultimately, the cost/benefit judgement must be made by the consumer.

What is the purpose and expected outcome of testing?

1. Decide on the purpose of testing and the intended use of the test results. Use the process of test selection as an opportunity to clarify the goals of the proposed assessment process. These goals should drive the rest of the selection process.

2. Decide what specific information is needed and the range of alternatives by which that information could be gathered (e.g., direct assessment, parent report, teacher observation).

3. How, specifically, will the information be used in the decision-making process? The consequences of testing for the child have important implications for the level of technical quality required.

What is the availability of high quality instruments appropriate for the children you will assess?

4. Examine the reviews and choose two or three instruments with appropriate content and age ranges to review in further detail. The reviews in this guide are by necessity brief, but more in-depth reviews are available from the sources listed in Appendix K.

5. Pay particular attention to evidence of reliability and validity appropriate to the specific test use.

6. Compare the normative population, if any, to the demographic characteristics of the population to be assessed.

How easily can you implement the use of this measure?

7. Consider the cost and the logistics for each instrument. Are the costs within available resources? (Include costs of obtaining the instrument, manual, test kit, consumable test forms, record sheets). What facilities or special equipment is needed? Is the time for administration reasonable?

8. Consider what training will be needed for administrators. Are training materials available?

9. Review the actual instruments and accompanying materials; either buy them or acquire them for examination through a test library. Review the test administration procedures with thought to the issues of training administrators, the appearance, sturdiness and cost of the kit and/or other materials, and the logistics of testing a large number of children.

References

Abidin, R. R., Golladay, W. M. & Howerton, A. L. (1971). Elementary school retention: An unjustifiable, discriminatory and noxious educational policy. *Journal of School Psychology*, 9, 410-417.

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

APA (1985). American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bredekamp, S. (Ed.) (1987). *Developmentally appropriate practice in early childhood education programs serving children from birth through age 8* (exp. ed.). Washington, DC: NAEYC.

California State Department of Education (1988). *Here they come: Ready or not.* Report of the School Readiness Task Force. Sacramento, CA: California State Board of Education.

Chew, A.L. & Morris, J.D. (1984). Validity of the Lollipop Test: A Diagnostic Screening Test of School Readiness. *Educational and Psychological Measurement, 44*, 987-991.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Fromberg, D. P. (1989). Kindergarten: Current circumstances affecting curriculum. *Teachers College Record,* 90(3), 392-403.

Goodwin, W. L. & Driscoll, L. A. (1980). *Handbook for measurement and evaluation in early childhood education.* San Francisco, CA: Jossey-Bass.

Hilliard, A. (1985). What is quality child care? In: B.M Caldwell and A.G. Hilliard III (Eds.), *What is quality child care?* Washington, DC: National Association for the Education of Young Children.

Joiner (1977). *A technical analysis of the variation in screening instruments and programs in New York State.* New York: City University of New York, New York Center for Advanced Study in Education, (ERIC Document Reproduction Service No. ED 154 596).

Kamii, C. (Ed.). (forthcoming, cited in Fromberg, 1989). *Achievement tests in early childhood education: Power in need of accountability.* Washington, DC: National Association for the Education of Young Children.

Meisels, S. J. (1985). *Developmental Screening in early childhood: A guide* (rev. ed.). Washington, DC: NAEYC.

Meisels, S. J. (1986). [National survey of early childhood special education policies and practices]. Unpublished raw data, cited in Meisels, 1987.

Meisels, S. J. (1987). Uses and abuses of developmental screening and school readiness testing. Young Children, 42, 4-9.

Meisels, S. J. and the Expert Team on Screening and Assessment, NCCIP (1988). Guidelines for the identification and assessment of young disabled and developmentally vulnerable children and their families. National Center for Clinical Infant Programs, National Early Childhood Technical Assistance System.

Michigan Department of Education (1984). Superintendents' Study Group on Early Childhood Education. Lansing, MI, cited in Meisels, 1987.

NAEYC (1985). Position statement on developmentally appropriate practice in early childhood programs serving children from birth to age 8. Washington, DC: NAEYC.

NAEYC (1988). Position statement on standardized testing of young children 3 through 8 years of age. Young Children, 43, 42-47. (Adopted November 1987)

NAESC/SDE (1987). Unacceptable trends in kindergarten entry and placement: A position statement of the NAESC/SDE. National Association of Early Childhood Specialists in State Departments of Education.

NASBE (1988). *Right from the start*. The report of the NASBE task force on Early Childhood Education. Alexandria, VA: National Association of State Boards of Education.

NBCDI (1987). *Safeguards: Guidelines for establishing programs for four-year-olds in the public schools*. Washington, DC: National Black Child Development Institute.

Peck, J.T., McCraig, G. & Sapp, M.E. (1988.) *Kindergarten policies: What is best for children?* Washington, DC: NAEYC.

Salvia, J. & Ysseldyke, J. E. (1988). *Assessment in special and remedial education (4th ed.)*. Boston, MA: Houghton Mifflin Company.

Sattler, J.M. (1988). *Assessment of Children, Third Edition*. San Diego, CA: Jerome M. Sattler, Publisher.

Shepard, L. & Smith, M.L. (Eds.), (In press). *Flunking grades: Research and policies on retention*. Policy Series in Education. New York: Palmer Press.

Shepard, L. & Smith, M.L. (1986). Synthesis of research on school readiness and kindergarten retention. *Educational Leadership, 44, 78-86.*

Shepard, L. & Smith, M.L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. The Elementary School Journal, 89(2), 135-145.

Teale, W. H. (1988). Developmentally appropriate assessment of reading and writing in the early childhood classroom. *The Elementary School Journal*, 89(2), 173-183.

World Book (1987). *Getting ready for school: What kindergarten teachers would like your child to know*. Chicago, IL: World Book, Inc.

APPENDIX A

SELECTION CHECKLIST FOR SCREENING INSTRUMENTS

Selection Checklist for Screening Instruments

I. Utility

A. Information Obtained

1. Is the stated use of this instrument to provide norm-referenced information on a broad range children's abilities in such a manner that it can be used to identify children at risk of potential learning problems?
2. Does the instrument provide scores which are easily calculated, readily interpreted, and useful for determining the refer/don't refer classification?
3. Does the manual provide information on developing local norms and cut scores?
4. Does the instrument cover the entire age range appropriately (i.e., no ceiling or floors in terms of scores within the age range to be screened)?
4. Does the instrument provide help with reporting to schools and communication with parents?
5. Is the instrument available and validated for the languages needed in your community?

B. Logistics

1. Is the instrument short and quick? How long does it take to administer?
2. Is the instrument easy to use? Who can administer the test (teachers, specialists, trained assistants), and what kind of training will be necessary?
3. Are training materials provided?
4. What kind of facilities and equipment are needed for administration?

C. Cost

1. Are the costs within available resources? Include costs of obtaining the instrument (manual, test kit, consumable test forms, record sheets, etc.), training administrators, and collecting data.

II. Validity

A. Evidence for Content Validity

1. Is the content appropriate to measure the broad range of underlying abilities that affect learning? How was the content determined in the test development process? Has the content been reviewed by experts?
2. Does the content completely cover what you intend to measure, or are there important areas not covered?

   Does the content cover skills in the following areas?:

   Language: receptive, expressive language skills
   Speech: articulation, fluency
   Cognition: reasoning, memory for objects or events in sequence
   Perception: visual, auditory discrimination
   Perceptual-motor, Fine, Gross motor coordination
   Personal/social skills, affect

Further content considerations include:

Does the instrument provide for parental input?

Does the instrument provide for vision, hearing, health, dental assessment?

3. Does the child understand what he/she is being asked? Is there evidence that the instructions, the format, and the response required are appropriate to measure what is intended, rather than attention span, cultural background or ability to speak English?
4. Will the screening experience be pleasant for young children?

B. Evidence for Criterion-Related Validity

1. Is there evidence of accuracy in classification; that is, are sensitivity and specificity at least 80 percent?
2. Is there other evidence that this measure predicts long-term outcome?
3. Is there evidence that this measure is related to other similar and valid measures?

C. Evidence for Construct Validity

1. Does the supporting material provide a definition of the aspects of children's abilities that it claims to measure? Does the test manual discuss, based on theory or research, how this definition was developed and why the test has the content it has?
2. Does performance improve with age, showing that the test measures developmental constructs?

III. Reliability

A. Is there evidence of *inter-rater* reliability?

B. Is there evidence of *stability* over time (*test-retest*)?

IV. Norms

A. Is the test norm-referenced?

B. Was the size of the norm group sufficient to have confidence in the norms (100/score grouping)? In particular, were there reasonable numbers of children in each age group?

C. How similar are the characteristics of the norm group (e.g., sex, race, geographic location, parental education) to the population which will be screened?

APPENDIX B

REVIEWS OF SCREENING INSTRUMENTS

Contents of Appendix B

Page		
1	**BSSI-S**	Basic School Skills Inventory - Screening
3	**BBCS-S**	Bracken Basic Concept Scale, Screening Forms
6		Brigance K & 1 Screen
9		Brigance Preschool Screen
11	**DASI II**	Developmental Activities Screening Inventory-II
13	**DIAL-R**	Developmental Indicators for the Assessment of Learning- Revised
17	**EISP**	Early Identification Screening Program
19	**ESI**	Early Screening Inventory
22	**FKSB**	Florida Kindergarten Screening Battery
24	**KLST**	Kindergarten Language Screening Test
26	**MAP**	Miller Assessment for Preschoolers
30	**PEER**	Pediatric Early Examination of Readiness

Instrument: Basic School Skills Inventory - Screening (BSSI-S, 1983)

Authors: Donald D. Hammill and James E. Leigh

Purpose: The authors intend this as a quick, easy measure to identify children who are "high risk" candidates for school failure.

Description: The **BSSI-S** consists of 20 items in a format which combines oral and performance responses with teacher ratings. It is designed to be used with children ages 4 to 6. The **BSSI-S** is individually administered and requires approximately 5-10 minutes to administer, depending on how well the administrator knows the child and the test. The "Answer and Record Sheet" contains instructions for administration and scoring, as well as the normative data tables. Responses are recorded on this same form, which also provides a chart for creating a profile of the standard scores for each subtest.

The **BSSI-S** covers skills across a broad scope of content as diverse as identifying coins, knowing the month of their birth and address, articulation of speech sounds, use of possessive nouns, counting, the ability to button, zip, snap and buckle.

Scoring: Items are scored on a pass/fail basis according to scoring criteria presented on the Answer and Record Sheet. For some items, the administration directions are standard and the scoring criteria are objective. For others, the teacher scores the item on the basis of knowledge or observations. In many items of this type, the scoring criteria are extremely subjective. Items that require the teacher's intepretation of terms like "appropriate" are particularly problematic on a norm-referenced test.

The child's raw score can be converted into standard scores and percentiles for each subtest and for the total, using tables on the record sheet.

Norms: Overall the norms are judged to be poor.

The standardization sample consisted of 376 children between the ages of 4-0 and 6-11 from 15 states. The sample was judged to be fairly representative of the characteristics of the U.S. population with regard to sex, race, and urban/rural residence. In terms of parent occupation, blue-collar workers were overrepresented (61% sample compared to 36% population), with a corresponding underrepresentation of white-collar workers. The sample was not representative in terms of regional distribution, with a substantial overrepresentation of Southern states.

The derived standard and percentile scores are based on the average scores of the standardization sample for each year of age (4, 5 and 6). Grouping by entire years is questionable during this period of rapid development. No information is presented as to the numbers of children tested at each age range and the mean and standard deviation of scores for the total sample. The median item difficulties (reported for a subsample of the standardization sample) of 81 and 85 suggest a serious ceiling effect. In fact, as the authors state, the "average" child in the six-year age group would pass 19 of the 20 items.

The ceiling effect does not necessarily mean that the test does not differentiate well among children at the lower end of the scale, which is the important area for a screening test. Children are judged to be "high risk" if their percentile score is 16 or below, corresponding to a standard score of 85, or one standard deviation below the mean. Considering ceiling effect in the context of the relatively small size of the normative sample and its skew toward less educated parents, one cannot place much confidence in the suggested "cut score."

As was mentioned, the lack of standard administration procedures and objective scoring criteria for many items makes norm-referenced interpretation questionable.

Reliability: Overall, the reliability is rated fair because of the lack of evidence for important types of reliability.

The author(s) present *internal consistency* reliability coefficients of .80 for age 4 and .83 for ages 5 and 6. *Alternate forms* reliability was examined on standardization scores between the BSSI-D and the shorter screening form, the **BSSI-S**. The correlations were .91, .92 and .88 for ages 4, 5 and 6, respectively. However, if both forms were scored by the same teacher at the same point in time, these correlations may represent a substantial overestimate of reliability.

Stability of measurement over time was not examined, nor was *inter-rater* reliability. As with the BSSI-D, because the **BSSI-S** is a measure of teachers' perceptions of children's abilities, the lack of evidence of inter-rater reliability is a serious issue.

Validity: Overall, the validity for the **BSSI-S** as a measure to identify children with potential learning problems is rated poor, because of the lack of information on the sensitivity and specificity of classifications based on BSSI-D results, and the dubious quality of the norms on which the classification cut score is based.

Content validity: The 20 items comprising the **BSSI-S** were selected from the BSSI-D on the basis of item discrimination and difficulty. No evidence is offered to justify the specific content. The problems with subjective administration and scoring criteria affect the content validity in terms of the appropriateness of the manner in which the content is measured.

Criterion-related validity: Concurrent validity of the **BSSI-S** was evaluated in relation to teacher ratings. The correlations of teacher ratings with the **BSSI-S** was .43. The value of this evidence is questionable since the ratings were for "general readiness" on a three-point scale, and the scores of the test were also largely teacher perceptions. The contention that the **BSSI-S** effectively measures the content of the BSSI-D was better supported. Correlations of the **BSSI-S** with the BSSI-D subtests ranged from .63 to .85; .92 with the BSSI-D total score.

Construct validity: General evidence is presented supporting the relationship between **BSSI-S** and chronological age. Evidence that the **BSSI-S** differentiates children diagnosed as "learning disabled" from "normal" children was presented for a sample of 12 children.

Utility: It is questionable whether the **BSSI-S** results add any value to teacher judgements on the identification of children with potential learning problems. The **BSSI-S** fails to meet most of the most important criteria to support its use for this purpose. The derived standard and percentile scores are based on the average scores of the standardization sample for each year of age (4, 5 and 6). Grouping by entire years is questionable during this period of rapid development.

Availability: Pro-Ed, 5341 Industrial Oaks Blvd., Austin, Texas, 78735.

Instrument: Bracken Basic Concept Scale - Screening (BBCS Screening, 1984)

Author: Bruce A. Bracken, Ph.D.

Purpose: The author's purpose is the provide a screening instrument to identify individual children in need of conceptual remediation or more intensive diagnostic evaluation.

Description: The **BBCS** Screening Tests (Forms A and B) each consist of 30 items, group administered in a paper and pencil, multiple choice format. The **BBCS** Screening Tests are intended for kindergarten and first grade children ages 5 through 7, and require approximately 15 minutes to administer. The manual for the **BBCS** Screening Tests is the same as for the **BBCS**.

The items are arranged, two to a page, in increasing order of difficulty in the test booklets. There are four answer choices per item, arranged 2 x 2. There is one practice item. The examiner must continually demonstrate where the children should be next in the test booklet and test proctors are needed to ensure that children are following the directions correctly.

Standard instructions for administration should be read exactly as printed in the manual. The child is instructed to "Put an X on the picture that ..." followed by an item stem that describes the correct answer. The test format necessitates lengthy instructions on turning pages, folding the book back and finding the correct item.

The basic concepts addressed by the **BBCS** have been grouped into eleven subtests. The items for the **BBCS** Screening Tests were drawn from eight of those subtests. The number of times chosen from each varies among subtests and between forms. The subtests and examples of item contents are listed as follows.

Comparisons: "boats that are **alike**", "boxes that are **not the same**"
Shapes: three-dimensional shapes, underlined, space
Direction/Position: outside, over, forward, right
Social/Emotional: old, difficult, exhausted, curious
Size: deep, large, medium-sized
Texture/Material: smooth, liquid, tight
Quantity: dime, neither, less than
Time/Sequence: starting, second, after

The manual presents administration and scoring procedures, suggestions for interpretation of results and instructional planning, as well as tables for scoring and technical information about the test development.

Scoring: All items are scored on a pass/fail (1/0) basis. Raw scores are converted into standard scores based on age in four-month intervals (5-0 through 7-0), or "concept ages" in one-month age intervals.

The standard scores can then be converted into percentile ranks, stanines, or normal curve equivalents (NCEs) by reference to a second table. Raw scores can also be converted in "concept ages" by one month age intervals (total score) or two month age intervals (subtest scores).

A group analysis form is available so that the teacher can look at conceptual performance for as many as 12 children at at time.

The author suggests that the "at risk" cutscores be fairly liberal in order not to miss children who need to be referred. It is suggested that any child who scores more than one standard deviation below the mean (a standard score of 85, at the 16 percentile) should be considered a candidate for more intensive evaluation. Alternatively, a cutscore of one standard deviation below the mean can be based on locally developed means and standard deviations. The manual presents an example of the development of a cutscore.

Norms: The norms are rated as fair.

The standardization sample consisted of 879 children, 559 in kindergarten and 320 in first grade. The standardization testing was conducted in small group sessions. All children were tested with both forms of the test, half taking Form A one day and Form B the next, half the reverse. All children were enrolled in public schools with a variety of ethnic groups and socioeconomic (SES) levels represented. No specific information is provided regarding SES levels.

The sample was somewhat representative of the 1980 US census distributions of sex, ethnic group and geographic region. Some of the demographic information is clearly presented for the full scale and diagnostic scale standardization samples combined. It is not clear if and how the screening and the diagnostic scale samples overlapped.

The sample was representative in terms of percentages by sex and ethnic group (black, white, hispanic, other). The southern and north central regions were under- and overrepresented by roughly 10%, respectively. The white and the "other" ethnic categories were under- and overrepresented by 10% No information is presented to assess the representativeness of the sample by age group.

There are 6 four-month age intervals used for translating raw scores to standard scores, and 26 one-month intervals used to translate raw scores into "concept ages". This would translate into more than 100 children per interval *if* the ages were evenly distributed. The numbers used to calculate the concept ages were much smaller. There are no tables in the manual that indicate the actual age distribution of the normative sample.

The percentage of children passing each individual item (item difficulties) are presented for screening forms A and B.

Reliability: The reliability of the **BBCS** Screening Tests is rated fair.

Stability over time was not addressed for individual forms. *Alternate forms* reliability ranged from .71 to .80. *Internal consistency* reliability coefficients ranged from .76 to .80.

Validity: Evidence for the validity of the **BBCS** Screening Tests is rated poor.

Content validity: The content validity of the **BBCS** Screening Tests is based on that of the **BBCS**. Items were selected for the two forms on the basis of item difficulty to match the forms in terms of difficulty level. The correlation between the two forms was only .51 for a sample of 47 kindergarten students, and .53 for a sample of 47 first grade students; surprisingly low for alternate forms.

After the standardization data was collected 5 of the original 35 the items were eliminated from each screening form.

Criterion-related validity: The only evidence for the validity of the **BBCS** Screening Tests is demonstrated through correlations of around .60 between the total scale and the separate forms of the screening tests for the samples of 47 kindergarten and 47 first grade students. The lack of evidence of predictive validity for the cutscores is a serious drawback.

Utility: The relationship between order of item difficulty of the items on the **BBCS** Screening Tests and the **BBCS** can be used to estimate performance on the diagnostic scale for instructional purposes.

There is not sufficient evidence of validity for the use of the **BBCS** Screening Tests to identify at-risk children. The relatively low correlations between the total **BBCS** and the Screening Tests may have been due to the group versus individual administration. This difference in administration may affect the at-risk students more than others. The effects of group versus individual administration should be examined.

As with the **BBCS**, there are a few items which are unnecessarily busy and might take longer for the child to visually isolate the information needed to understand the concept.

The use of concept ages is particularly problematic with the sample because there were so few children for each age interval. As with the **BBCS**, it is conceivable that some children in the normative sample may have been in school one year longer than children of exactly the same age. The concept age averages the performance of these children.

Availability: The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204

Instrument: Brigance K & 1 Screen (1982)

Author: Albert H. Brigance

Purpose: The authors' purpose is to provide a screening instrument to assist in program planning and to identify children in need of more intensive diagnostic evaluation for potential learning problems, to determine appropriate placement and to assist in program planning.

Description: The **K & 1 Screen** includes 12 skill areas for kindergarten and 13 for grade 1. The format requires a variety of oral, pointing, performance and motor responses. The **K & 1 Screen** is individually administered and requires approximately 10 to 15 minutes to administer and score.

The test consists of a spiral bound book that contains directions for administration and scoring as well as stimulus pictures for some items. The clearly written test manual also provides a discussion of test development, general instructions for setting up screening stations, and suggestions for interpretation of results.

Responses are recorded on separate data sheets for kindergarten and grade 1 children. These data sheets are conveniently formated with many cues for test administration and scoring.

The **K & 1 Screen** covers the areas described below in separate subtests of a few items each. The items for grade 1 represent an upward extension in terms of difficulty of the items for kindergarten.

Kindergarten	*Grade 1*
Personal data response	Personal data response
Color recognition	Color recognition
Picture Vocabulary	Picture Vocabulary
Visual Discrimination-A	Visual Discrimination-B
Visual-Motor Skills	Visual-Motor Skills
Gross Motor skills	Gross Motor skills
Rote Counting	Rote Counting
Identification of Body Parts	Draws a Person
Follows Verbal Direction	Recites Alphabet
Numerical Comprehension	Numerical Comprehension
	Numerals in Sequence
Prints Personal Data	Prints Personal Data
	Recognition of Lowercase Letters
	Recognition of Lowercase Letters
Syntax and Fluency	Auditory Discrimination

There is a space on the answer sheet to record observations of such things as handedness, pencil grip, and speech quality. A screening observations form is provided for the examiner to record any observations of specific problems with vision, hearing, speech, self-reliance, emotional function, motor skills or physical appearance (health).

A separate teacher's rating form, echoed by a parent rating form, has 38 questions evaluating children's behavior according to such criteria as demonstrating number and verbal concepts, self helf, social and motor skills.

Scoring: Scoring criteria and examples of scoring are presented with the administration directions for each item. A point value can be assigned to each item (as indicated above) and summed to provide a possible total score of 100. The author suggests the use of these scores for ranking children and determining referral cutscores.

Norms: The **K & 1 Screen** is not normed. Although it was field-tested, no data is presented in the manual.

The author provides a procedure for creating locally relevant cutscores by ranking children into categories on the basis of total scores. This could lead to quite a variable basis for referral or placement, dependent on the size and nature of the group being tested at any given time. A more rigorous method for developing local norms could be provided using means and standard deviations and collecting data over many groups.

There is a place on the child's data sheet to record whether the total score was lower, average, or higher than the group tested (on the basis on dividing the sample into groups). This could be an extremely misleading piece of information to have on a child's record when there is nothing to indicate the nature of the group with which the child was compared. [There is a place to indicate whether the child was younger or older than the other members of the group, but not in absolute terms or by what magnitude.] The use of the term "average" is also inappropriate given the method of ranking. Depending on the distribution of scores, differences between children in different groups may not have any practical significance.

Reliability: No evidence is provided for the reliability of the **K & 1 Screen**. This is not acceptable for a measure that is used to guide important decisions such as referral and placement.

Validity: Evidence for the validity of the **K & 1 Screen** is rated good, primarily on the basis of a separately published research study, but is limited.

Content validity: The items for the **K & 1 Screen** were selected from the **Brigance Inventory of Early Development** and the **Inventory of Basic Skills**. Items were selected on the basis of predictive validity for success in school (as indicated by the research literature), feasibility, objectivity, field recommendations of appropriateness, and insurance of success experiences for the child screened. The **K & 1 Screen** was extensively field-tested in 53 schools in 14 states. A summary of the percentage of raters who viewed skills areas as appropriate is presented in the manual.

Criterion-related validity: The author does not present evidence of concurrent or predictive validity, however there is at least one published study that strongly supports the predictive validity of the **K & 1 Screen**. Gordon (1988) administered 20 subtests of the **Inventory of Basic Skills** , virtually identical to the content of the **K & 1 Screen**, to 109 beginning kindergarten children. The children were tested with the **Stanford Achievement Test** (**SAT**) in second grade. A classification analysis using the **SAT** score which would make children eligible for Chapter I services as a "failure" criteria, the total score of the **K & 1 Screen** had a sensitivity (correct referral) of .90, a specificity (correct no-referral) of .76 and an overall hit rate of .80.

[Gordon, R. R. (1988). Increasing efficiency and effectiveness in predicting second-grade achievement using a kindergarten screening battery. *Journal of Educational Research*, Volume 81(no. 4), 238-244.]

Utility: The **Brigance K & 1 Screen** was developed on the basis of requests from users of the Brigance inventories. It is an attractively presented and easily administered test. Evidence presented by the author does not meet the most important validity and reliability requirements for a screening instrument. In addition, the author recommends some questionable practices in terms of the development and use of cutscores.

There is strong but limited evidence from one study that total scores from the items on the **K & 1 Screen** identify children who later qualify for referral for special services with high degrees of sensitivity and specificity. At the present time, if the **K & 1 Screen** is to be used for screening, the user must take the time to establish local cutscores with adequate validity and reliability. This is a burden that the test developer should take on if this instrument is to be marketed as a screening device.

Availability: Curriculum Associates, Inc., 5 Esquire Road, North Billerica, MA 01862-2589.

Instrument: Brigance Preschool Screen for Three- and Four-Year-Old Children (1985)

Author: Albert H. Brigance

Purpose: The authors' purpose is to provide a screening instrument to identify children in need of more intensive diagnostic evaluation for potential learning problems, to determine appropriate placement and to assist in program planning.

Description: The **Preschool Screen** includes 44 items for three-year-olds and 46 items for four-year olds in a format which combines oral, pointing, performance and motor responses. The **Preschool Screen** is individually administered and requires approximately 10 to 15 minutes to administer and score.

The test consists of a spiral bound book that contains directions for administration and scoring as well as stimulus pictures for some items. The clearly written test manual also provides a discussion of test development, general instructions for setting up screening stations, and suggestions for interpretation of results.

Responses are recorded on separate data sheets for three- and four-year old children. These data sheets are conveniently formated with many cues for test administration and scoring.

The **Preschool Screen** covers the 11 areas described below in separate subtests of a few items each. The items for four-year-olds represent an upward extension in terms of difficulty of the items for three-year-olds. The three different tasks for four-year-olds are indicated in parentheses. The table indicates the number of items and the item weights for scoring.

No. Items x Item Weight

Age 3	Age 4	Task
3 x 2	4 x 1	Personal data
9 x 1	9 x 1	Identify body parts
3 x 3	3 x 3	Gross motor skills
3 x 3	3 x 3	Identifies object (Tells use of objects)
3 x 3	3 x 3	Repeats sentences
3 x 3	3 x 3	Visual-motor skills
3 x 3	3 x 3	Number concepts
5 x 2	5 x 2	Build tower with blocks
5 x 2	5 x 2	Matches colors (Identifies colors)
5 x 2	6 x 2	Picture vocabulary
2 x 5	2 x 5	Plural *s* and *-ing* (Prepositions and irregular plural nouns)

There is a space on the answer sheet to record observations of such things as handedness, pencil grip, and speech quality. A screening observations form is provided for the examiner to record any observations of specific problems with vision, hearing, speech, self-reliance, emotional function, motor skills or physical appearance (health).

A separate teacher's rating form, echoed by a parent rating form, has 38 questions evaluating children's behavior according to such criteria as demonstrating number and verbal concepts, self helf, social and motor skills.

Scoring: Scoring criteria and examples of scoring are presented with the administration directions for each item. A point value can be assigned to each item (as indicated above) and summed to provide a possible total score of 100. The author suggests the use of these scores for ranking children and determining referral cutscores.

The author also suggests that for referral purposes testing can be stopped once the child has enough points to pass a pre-established cutscore. This may mean, however, that the child would not be tested on most of the language items, and a deficit in this area would be missed.

Norms: The **Preschool Screen** is not normed.

The author provides a procedure for creating locally relevant cutscores by ranking children on the basis of total scores and referring those in the lower third. This could lead to quite a variable basis for referral, dependent on the size and nature of the group being tested at any given time. A more rigorous method for developing local norms could be provided using means and standard deviations and collecting data over many groups.

There is a place on the child's data sheet to record whether the total score was lower, average, or higher than the group tested (on the basis on dividing the sample into three groups). This could be an extremely misleading piece of information to have on a child's record when there is nothing to indicate the nature of the group with which the child was compared. [There is a place to indicate whether the child was younger or older than the other members of the group, but not in absolute terms or by what magnitude.] The use of the term "average" is also inappropriate given the method of ranking. Depending on the distribution of scores, child in the lower third could have a score very close to a child in the upper third.

In general the author recommends that children scoring 60 or below be referred for more intensive evaluation. No rationale is presented for this number.

Reliability: No evidence is provided for the reliability of the **Preschool Screen**. This is not acceptable for a measure that is used to guide important decisions such as referral.

Validity: Evidence for the validity of the **Preschool Screen** is rated fair and is based entirely on content validity.

Content validity: The items for the **Preschool Screen** were selected from the **Inventory of Early Development**. A field-test edition was reviewed by early childhood educators, administrators, consultants, psychologists and special education teachers from 12 states. The field test sample is not described.

Utility: The Brigance **Preschool Screen** was developed on the basis of requests from users of the *Inventory of Early Development*. It is an attractively presented and easily administered test. However it does not meet the most important validity and reliability requirements for a screening instrument. In addition, the author recommends some questionable practices in terms of the development and use of cutscores. If the **Preschool Screen** is to be used for screening, the user must take the time to establish local cutscores with adequate validity and reliability.

Availability: Curriculum Associates, Inc., 5 Esquire Road, North Billerica, MA 01862-2589.

Instrument: Developmental Activities Screening Inventory (DASI-II, 1984)

Authors: Rebecca R. Fewell and Mary Beth Langley

Purpose: The authors' purpose is to provide an informal screening measure for children functioning between the ages of birth to five. The test was designed to be easily administered by classroom teachers and be directly applicable to the content of a child's preschool or home-based program. It was specifically designed to be non-verbal so that it does not penalize children with auditory impairment or language disorders. Adaptations for administering the test to visually impaired children are clearly specified in the manual.

Description: The **DASI-II** consists of 67 items, in a primarily performance response format. There are six items for each of 11 levels (approximately six month age intervals from birth to age 5-0). The test covers the following 15 functions (most items fall into multiple categories, included are examples from levels appropriate for ages 3-5):

Sensory intactness: identifies colors, copies bead patterns
Sensorimotor organization: matches blocks to set configuration, copies circle, cross
Visual pursuit/object permanence: (younger age levels)
Means-ends relationships: (younger age levels)
Causality: (younger age levels)
Imitation: imitates diagonal paper fold
Behaviors relating to objects: (younger age levels)
Construction of objects in space: (younger age levels)
Memory: follows two step command, identifies colored blocks from memory
Discrimination: names colors, stacks rings in correct order
Association: matches pairs of pictures to indicate functional associations
Quantitative reasoning: understands concepts of two and three
Seriation: stacks five rings in order by size, copies bead patterns
Spatial relationships: copies forms, imititates diagonal paper folds, builds pyramid of six blocks
Reasoning: classifies pictures into three groups

The **DASI-II** is individually administered; no time requirements are noted in the manual, perhaps because of the wide range of ages covered. The examiner is advised to begin testing one level below his or her estimate of the child's developmental age. Ease of administration was a primary goal for the authors and the procedures are described "clearly, simply, and in non-technical language." Stimulus cards with pictures, shapes, words and numerals are included as part of the test package; other materials commonly present in preschool settings. (e.g., blocks) need to be assembled by the examiner.

Scoring: The manual presents scoring criteria after each item. The raw score, the sum of all items answered correctly, is converted into a developmental age (in months) using a table provided. The developmental quotient is computed using this developmental age. A rough guide to interpreting the significance of the Developmental Quotient ("Superior" to "Poor") is also provided. There is no explanation of how the raw scores corresponding to each developmental age were determined.

The manual includes only very general interpretation guidelines and instructions for teaching the skills addressed by the **DASI-II**.

Norms: Norms have not been established. No descriptive statistics, such as means, medians, standard errors of the mean and standard deviations, are presented in the manual.

Reliability: No data on reliability are provided in the manual.

Validity: Overall evidence of the validity of the **DASI-II** is rated poor.

Content validity: The **DASI-II** is a revised version of the original DASI, differing from the original in the addition of two levels at the lowest age range and in the replacement of three other items. The manual offers more justification of the appropriateness of the test format than the test content. Face validity and user feedback appear to have been the primary determining factors in item selection.

The **DASI-II** covers skills that "represent behaviors frequently included in tests of early cognitive development." The authors noted that basic materials such as paper, markers, blocks and beads were already present in preschool settings and that items on preschool assessment measures were similar to the tasks being taught using such materials.

The manual states that the test was designed to be non-verbal; however there are a few items which require a verbal response. There is no mention of any examination of the comparability of items administered verbally versus with gestures, or of verbal versus alternative response formats.

Construct validity:. The authors obtained a strong correlation between scores on the original DASI and scores on either the *Infant Intelligence Scale* (Cattell, 1940) or the *Merrill-Palmer Scale of Mental Tests* (Stutsman, 1948) for a sample of children known to have multiple disabilities. The age range of the 45 children is not specified. No relationship was found between the DASI and language measures, supporting its non-verbal nature.

Criterion-related validity. Evidence of concurrent validity is presented only for the original DASI. Without an empirical comparison of the tests it is hard to say how much such evidence can be generalized to the **DASI-II**. A strong correlation was found (for a very small sample) between the DASI and the *Developmental Assessment of the Severely Handicapped*. For two separate samples of delayed and non-delayed children (42 children ages 0-7 to 6-2, 14 "day-care" children ages 1-3 to 4-8) the DASI was strongly related to the *Preschool Attainment Record* (.97, .92) and the *Denver Developmental Screening Test* (.95, .87).

Utility: The **DASI-II** is a brief, easily administered test designed to be almost entirely non-verbal in response format. Although it is designed to be used for screening, there is no evidence of reliability and extremely limited evidence for validity. There are no data supporting the use of the Developmental Age scores or Developmental Quotients. The manual offers suggestions for teaching the concepts addressed in the **DASI-II** during the interim between initial identification and a comprehensive diagnostic assessment. This is a questionable practice which could lead to less accurate diagnosis.

Availability: PRO-ED, 5341 Industrial Oaks Blvd., Austin, Texas 78735

Instrument: Developmental Indicators for the Assessment of Learning - Revised (DIAL-R, 1983)

Author: Carol D. Mardell-Czudnowski, Ph.D. and Dorothea S. Goldenberg, Ed.D.

Purpose: The authors' purpose is to provide a screening instrument to identify preschool children in need of more intensive diagnostic evaluation for potential learning problems or for giftedness.

Description: The **DIAL-R** includes 24 items in a format which combines verbal and a variety of performance responses (e.g, catching, building, drawing). It is designed for children ages 2 years to 6 years. The **DIAL-R** is individually administered and requires approximately 20-30 minutes to administer and score. The format is designed so that administration can be done by teams, with different examiners administering the Motor, Language and Concepts parts of the test.

Many **DIAL-R** items are administered with the use of large dials which are mounted on stands and provide the stimulus pictures. Other than a watch with a second hand, materials for the photographs (instant camera, film and flashbulbs), nametags and warm-up activities (clay), everything needed for administration is included with the kit.

For each item the manual describes the materials needed, the procedure for administration, including standard instructions which should be read exactly as printed and detailed criteria for scoring. At the end of each section there is a list of eight behaviors which could influence testing (e.g., distractible, cries/whines). Occurrence of any of these behaviors is recorded at each testing station by circling a number (1-8) on the score sheet. Separate administration booklets for each area (Motor, Concepts, Language) are also provided.

The manual includes general directions for administration, scoring and interpretation of results as well as cautions about the appropriate and inappropriate use of the test and results. The manual also provides a discussion of test development and technical quality.

The **DIAL-R** content can be organized into three general screening areas. The following are examples of specific item content for each area.

Motor	Catching, jumping, hopping, skipping Building, cutting, matching, writing name
Concepts	Name colors, letters, counting, sorting
Language	Articulating, naming nouns and verbs, classifying Giving personal data, problem solving

Age related entry levels and exits are marked for each area in order to pace the administration for children of different ages and abilities. A parent information form is also available.

Scoring: The child's responses are recorded on individual scores sheets, including copying figures. Items are scored on the score sheet according to detailed scoring criteria presented with the administration directions for each item. If the child corrects an error without assistance, the best score is recorded. Raw scores for each item are then converted into scaled scores conveniently indicated on the score sheet. The area score is the sum of the scale scores for the eight items. All three areas are summed to obtain the total score.

The total score identifies the child as "Potential Problem", "OK" or "Potential Advanced" on the basis of cutscores determined for each three-month age group. The extreme category cutscores correspond to 1.5 standard deviations below and above the mean for each age group, the highest and lowest 6.68 percent. Between -1.5 SD and +1.5 SD is considered "OK". The primary cutscores are based on a sample representative of the US population in terms of race, as described below. Different cutscores are available for all white populations, all nonwhite populations as well as for the 5th and 95th percentiles and the 10th and 90th percentiles of the racially balanced subsample.

Because research indicated that the total score may overidentify 'potentially advanced' students, or mask a potential problem, cutscores were also developed by area. The authors urge the user to determine the most appropriate comparison group and cutscores for their particular population.

Percentiles based on total score (in intervals of 5) by six-month age group are available for the total sample, the all-white, and all nonwhite subsamples. The manual also gives a cutscore for the number of problem behaviors circled for each year of age, suggesting that higher numbers of problems behaviors merit a referral for social/affective problems.

Norms: Overall the norms are judged to be fair.

The 1983 standardization sample consisted of 2447 children between the ages of 2-0 and 5-11. Children were oversampled for the nonwhite category so that separate norms by race and age group could be established.

The sample was judged to be fairly representative of the characteristics of the U.S. population with regard to sex and geographic region, with a slight overrepresentation of the South at the expense of the West. With regard to community size, there were only eight primary testing sites (representing six states), which were approximately equally divided with populations above and below 50,000.

Information on socioeconomic (SES) level was not collected on all students and is not presented in the manual. The manual does present correlations between parental education level and **DIAL-R** total scores from the subsample for which this information was available. These correlations are statistically significant (higher parental education related to higher scores), but only moderate in size (.22 to .35). Approximately seven percent of the sample came from homes where a language other than English was spoken regularly.

The total sample was 55.5% white and 44.5% nonwhite (Black, Native American, Alaskan Natives, Asian, Pacific Islander and Hispanics of nonwhite racial background). A subsample of 1861 children was selected to be representative of the 1980 Census figures (73 % white, 27% nonwhite) and this subsample was used to determine the primary cutscores.

The cutscores are determined for 3 month age intervals. The number of children in each age interval is not reported for any of the samples used to determine cutscores. Judging from the distribution of children in the unreduced sample, there were probably fewer than 100 children in some age categories. The normative information for children over 6-0 is based on extrapolation because data was not collected above this age. Its use is not recommended.

Separate cutscores were determined for the total white and the total nonwhite populations and are presented in appendices in the manual. The racial composition of the nonwhite population is not described. It would appear that there were less than 100 per age group in the all white and all nonwhite subsamples. 'It is clear from a comparison of cutscores that the performance of the white and nonwhite samples was very different, however the means and other descriptive data are not provided. It would be informative to be able compare the level of parental education on these two samples.

The nonwhite norms are also hard to interpret because the distribution of minorities and the areas of the country where they live are not described. A population that was largely Asian from Hawaii would not be an appropriate "minority" reference group for a largely black Head Start program on the East coast!

Reliability: Evidence for the reliability of the **DIAL-R** is rated fair.

The authors present evidence of three types of reliability. *Stability* of measurement over time (*test-retest*) was assessed with a sample of 65 children (14-18 from each yearly age group), selected from the standardization sample. The correlation for the total score was .87, with correlations for Motor, Concepts and Language areas .76, .89 and .77, respectively. These correlations are somewhat inflated because of the range of ages included. No stability information is provided for the cutscores.

Internal consistency reliability was estimated on the basis of total and area scores. The overall coefficient was .96. These were calculated separately by age level and range from .75 to .94 for the total score, and from .41 to .88 within separate areas. The reliablity evidence does not support the use of area scores. It is problematic that the authors indicate that overreferrals of advanced and underreferral of problem children have been reported using the total score, since that is the only reliable score.

Evidence of *inter-rater* reliability is not presented.

Validity: Overall, evidence for the validity of the **DIAL-R** is rated fair.

Content validity: The **DIAL-R** is a revision of the the **DIAL**, published in 1972. 21 of the 24 items are unchanged or revisions of **DIAL** items. Evidence of the validity of the **DIAL** is presented to support the validity of the **DIAL-R**. Tasks were selected on the basis of teacher input to reflect behavior expected of children in the prekindergarten and kindergarten age range. Each task was also reviewed by professors in various fields related to early childhood education. A collection of behavioral 315 tasks was reduced to 155 on the basis of logic and further reduced on the basis of pilot studies. These 155 were clustered into 32 items for the standardization edition.

Criterion-related validity: **The Stanford-Binet Intelligence Scale** was chosen as a criteria to assess concurrent validity because it covers the same content areas as the **DIAL-R**, across the entire age range. The **DIAL-R** is not an intelligence test, but both tests should be related to school success. Correlations between the two tests were .40 for the total score, and .28, .50, and .33 for Motor, Concepts and Language areas, respectively. Classification analysis was also carried out using the **DIAL-R** cut scores. In terms of screening for just the potential problem end of the classification, the **DIAL-R** showed a sensitivity of 64% (correct referral), and a specificity of 97% (correct no-referral). The rate of underreferral was just 2% and all three of the children underreferred were 3 years of age or younger.

Several studies to assess predictive validity of the **DIAL-R** were underway at the time of publication. A longitudinal study of the predictive validity of the original **DIAL** found significant relationships between **DIAL** scores and achievement tests in kindergarten and first grade. A number of other studies conducted on the **DIAL** over the past decade are mentioned but the results are not summarized.

Construct validity: The validity of **DIAL-R** as a measure of developmental trends was examined, with an aggregate correlation between **DIAL-R** total score and age of .98. A factor analysis which resulted in only two factors, Motor and Concepts combined and Language does not lend support to the use of separate area scores.

Utility: There are many aspects about the **DIAL-R** that make it an appealing choice as a screening instrument, however the evidence of technical quality is marginal in term of making important educational decisions. The **DIAL-R** is an attractively and conveniently packaged, easily administered screening instrument. The plan for setting up screening stations and the roles of the screening participants are very well presented. The effort the authors put into creating special norm groups is commendable, however more detailed information is needed to determine appropriateness for individual screening sites. There is not sufficient evidence to support the reliability and validity of area scores and more studies need to be conducted to determine the validity of the total cutscores. There is no technical evidence to support the social/affective ratings.

The **DIAL-LOG**, a microcomputer-based system for scoring, reporting and record keeping can be used in the the development of local norms and cutscores. A trainng videotape is available, as well as a packet of test results and role playing activities. The **DIAL-R** Activity Card System provides school and home follow-up instructional activities keyed to the **DIAL-R** tasks. However, use of these might constitute "teaching to the test" and invalidate repeated screenings with the **DIAL-R**.

Availability: Childcraft Education Corporation, 20 Kilmer Road, Edison, New Jersey 08818

Instrument: Early Identification Screening Program (EISP, 1982)

Authors: The **EISP** was developed by the Baltimore City Public Schools, Office of Continuum Services, Division for Exceptional Children.

Purpose: The authors' purpose is to provide a measure for screening at the beginning of kindergarten and grade 1. The test covers auditory, visual, and expressive skills. The results can be used as a measure of skill development in these areas for instructional planning, as well as to identify children at risk of potential learning problems, and assist in documenting the need for referral and planning further evaluation.

Description: The **EISP** has one form with separate levels for kindergarten and grade 1. Each level includes three subtests which consist of one activity that addresses a combination of skills as follows:

Hear-Write: draw a series of figures (numbers for grade 1) named in succession by the examiner [taps auditory discrimination, short-term memory, beginning penmanship, and fine muscle control]

See-Write: copy a series of figures presented in the Screening Booklet (letters for grade 1) [taps visual discrimination, eye-hand coordination, beginning penmanship, and fine muscle control]

See-Say: name and point to colors on the See-Say Colors Chart (six colors; letters for grade 1) [taps general information, verbal skills, reading readiness, eye-hand coordination, and articulation]

The **EISP** is individually administered in three sessions on three consecutive days, requiring a total of 20 minutes for administration on all three days (8-10 on the first day because of practice items, 5 on subsequent days). The manual provides exact wording for the administrator. The child must respond on each task (drawing from verbal and visual stimuli, naming stimuli) as quickly as possible. Each activity is presented at each of the three screening sessions.

No specific training is required to administer the **EISP**; however some practice is needed to ensure accurate timing and counting of responses, particularly on the See-Say activity, as there is no provision for recording responses on the record form.

Materials include the Administration and Scoring Manual, consumable Screening Booklets for kindergarten and grade 1, two See-Say charts, a Class Record Sheet and a Ranking Worksheet. The manual provides instructions and helpful suggestions on setting up screening activities to minimize disruption of normal classroom activities.

Scoring: The test is timed, the child is allowed one minute for each of three attempts previously described. Scores for each subtest are the frequency of correct responses per minute, averaged across the three sessions. If the child misses one session, an average of two is allowed. Only limited scoring guidelines are presented, which may be problematic for inexperienced administrators. Groups of administrators are encouraged to develop their own scoring standards.

A local comparison group is created by ranking the average score on the three administrations of each subtest and a total across subtests for each classroom or grade level. The manual suggests that the lowest 25 % of scores could be considered a cut-off criterion for some form of intervention, however this depends of the nature of the population being tested and the nature of the decision being made.

Norms: The **EISP** is not normed.

Reliability: Evidence of reliability of the **EISP** is rated good.

Test-retest reliability was examined in a validation study of approximately 124 children selected randomly from a total of 558 kindergarten and grade 1 children from four schools in one large urban school district. Reliability coefficients ranged from .90 to .92.

Validity: Evidence of the validity of the **EISP** is rated fair.

Content validity. Items were were chosen on the basis of their relevancy to expected classroom performance and on the basis of observation of student performance. The three activities for the final version were selected from a pool of nine, on the basis of a pilot study. No theoretical or empirical rationale for item selection is presented. No rationale or data are presented supporting the one-minute time limit or the three administrations of each activity. Presumably repeated administrations ensure stability of the results; there is a caution that the use of only one session would result in a large margin of error.

Criterion-related validity. The validation study examined concurrent validity by the relationship of scores on the **EISP** to teacher ratings and to with scores on the language and math subtests of the **Test of Basic Experience (TOBE)**. The correlations with **TOBE** were low (.37 and .33, for grades K and 1, respectively). There was 89% agreement of children identified by teachers (at the beginning of the year) as at-risk and low-risk. Predictive validity was supported (for a separate group of children) by a 93% agreement with teacher ratings of children at the end of the year. The contrast between low correlations with the **TOBE** and high agreement with teacher ratings suggests that something other than academic aptitude is being measured by the **EISP**.

Of the 124 child sample for the validation study, 55 were in kindergarten and 69 were in grade 1. The small size of the sample and the fact that 96% were black limit the generalizability of the results to the general school population.

Utility: The **EISP** is a quick, easily administered assessment of some of the skills required in the classroom. However, its utility as a screening instrument has not been established. Administrators must take the time to work out clear scoring guidelines if consistent scoring is to be provided.

The time limits may be frustrating to some children. The issue of color-blindness is not addressed but should be kept in mind as a possible explanation for problems on the See-Say task (kindergarten level). There is a Spanish-directions supplement to the manual but no separate technical information is provided.

Availability: Modern Curriculum Press, 13900 Prospect Road, Cleveland, Ohio 44136 (216-238-2222).

Instrument: Early Screening Inventory (ESI, 1983)

Authors: Samuel J. Meisels and Martha Stone Wiske

Purpose: The authors intend this as a brief, easily administered, developmental screening instrument to identify children who are in need of further diagnostic evaluation.

Description: The **ESI** consists of 30 items in a format which combines oral and performance (counting, building, drawing, movement) responses. It is designed to be used with children ages 4 to 6. The **ESI** is individually administered, requiring approximately 15-20 minutes for each child. The manual contains standard instructions for administration, directions for scoring and interpretation as well as technical information. Instructions to the child should be read exactly as they are written in the manual. A score sheet is used to record and score children's responses. The exact wording of instructions and information about prompts is conveniently printed above each item on the score sheet. Space is also provided next each item for examiner's comments.

The **ESI** is a brief survey of development across a broad range abilities including speech, language, cognition, perception, and gross and fine motor coordination. It is divided into four sections, the last three representing general areas of development.

Initial screening items:
- A. Draw a person (scored); write name or letters (unscored)

Visual-Motor/Adaptive:
- A. Copy forms (circle, cross, square, triangle)
- B. Visual sequential memory (placement of three forms)
- C. Block building

Language and Cognition:
- A. Number concept (counting, altogether)
- B. Verbal expression (child's ability to name and tell about color [red, yellow, blue, green], shape, use and other attributes of a ball, toy car, wooden cube and button)
- C. Verbal reasoning (opposite analogies, e.g., brother is a boy: sister is a _____)
- D. Auditory sequential memory (3 and 4 digits)

Gross Motor/Body Awareness:
- A. Balance
- B. Imitate movements (arms)
- C. Hop
- D. Skip

Other information recorded (but not scored) includes color matching (if the child does not identify all colors in the verbal expression item), speech errors (consonent, vowel, intelligibility), other language errors and use of complete sentences.

A Parent Questionnaire which accompanies the **ESI** is not scored. It provides a context for the results of the screening test in terms of family, health and developmental risk indicators. The four sections include *basic information* (parents' educational level, the family configuration, child's educational experience), the child's *medical history*, *health*, and *development* (temperament and developmental milestones). The questionnaire may take as long as 15 minutes to complete. If it is administered before the screening it can cue the examiner to look for specific difficulties. The questionnaire can also be used to interpret the results of the screening to the parent.

Scoring: Every item on the **ESI** is administered. Items are recorded as "pass", "fail" or "refuse" on the basis of scoring criteria presented in the manual, following the instructions for administration of each item . After administration is completed, the number of points received for each "pass" can be calculated. This number ranges from 1 to 3. Most items can be scored quickly and easily. However, the inexperienced examiner may spend more time scoring the "copy forms" and "verbal expression" items. Scoring criteria for the "verbal expression" items are somewhat confusing, particularly in terms what should be credited as "other" attributes.

The child's raw score is converted into "OK" , "Rescreen" or "Refer" recommendation categories using **ESI** norm-based cutscores or locally developed cutscores. The **ESI** cutscores are based the norms described below and represent one standard deviation (rescreen) and two standard deviations (refer) below the mean for a given six-month age interval (4-0 to 4-5, 4-6 to 4-11, 5-0 to 5-5, and 5-6 to 5-11). Children who score in the "Rescreen" range should have the **ESI** readministered in 8 to 10 weeks, unless there is some other indication that further evaluation should be done immediately.

The total score of the **ESI** is used to determine whether to refer or rescreen. Because each ability is sampled with only a few items scores on any one ability or domain should not be interpreted to reflect general ability in that area.

Norms: The normative information is rated fair. However, the **ESI** is in the process of being re-normed with a representative, national sample in both English and Spanish. The new standardization should alleviate any reservations about use of the norms.

The standardization sample consisted of 465 children between the ages of 4-2 and 5-10. The sample characteristics are not well described, other than that is consisted primarily of Caucasian children from low to lower-middle socioeconomic status urban families. The manual recommends that those using the **ESI** on a large scale establish their own cutscores using one and two standard deviations below the mean of the local scores.

There were reasonable numbers of children in the 4-6 to 4-11 and 5-0 to 5-5 age ranges, but only 50 younger and 13 older. The cutscores for the older age range were based on extrapolations of data from younger children.

Reliability: The reliability of the **ESI** is rated good but limited because it is based on small samples of children and the correlations cover a wider age range than the score intervals.

The authors present evidence of *inter-rater* reliability, ranging from .80 and higher for subtests to .91 for the total score. *Stability* of the over time *(Test-retest* reliability) was demonstrated with a correlation of .82 for the total score, although the correlations for the subtest scores were all below .80. No evidence is presented concerning the reliability of the cutscore categories (OK, Rescreen, Refer).

Validity: Evidence for the validity of the **ESI** is rated good.

Content validity: The content of the **ESI** is based on well-known and widely used developmental tests. In fact, several items are attributed directly to the **Illinois Test of Psycholinguistic Abilities**, the **Stanford-Binet**, the **Denver Developmental Screening Test**, and the **Purdue Perceptual-Motor Survey**. The **ESI** underwent four major revisions based on field tests with more than 3000 children. An analysis contrasting the OK and Refer groups indicated that almost all items clearly discriminated between these groups.

Criterion-related validity: Evidence of concurrent validity was established by comparing results of the **ESI** with the **McCarthy Scales of Children's Abilities (MSCA)** for a stratified sample of 102, primarily caucasian children from the metropolitan Boston area. A correlation of .73 was obtained between scores for the two tests. A classification analysis was done using the **ESI** OK, Rescreen and Refer categories. Categories for the **MSCA** were calculated on the same basis as the **ESI** cutscores (one and two standard deviations below the mean) showed strong agreement on outcome with an overall hit rate of 89%, a sensitivity of 87% and a specificity of 90%.

A similar study was done to provide evidence of short-term predictive validity using the **Metropolitan Readiness Test** (**MRT**) as a criterion measure. A group of 472 children were screened with the **ESI** before kindergarten and tested with the **MRT** at the end of kindergarten. Correlations between the **ESI** and the **MRT** ranged from .44 to .49 across age and sex groups. A classification analysis using the 15th percentile as a cutscore for both measures showed an overall agreement of 83% with a sensitivity of 33% and specificity of 91%. Classification analyses for a sample of 115 children followed through grade 4, using report card grades as the criterion measure, showed sensitivities ranging from 100% (grade 2) to 50% (grade 4) and specificities ranging from 82% (grade K) to 61% (grade 3). Using the 15th percentile cutscore, approximately one standard deviation below the mean, the **ESI** is more likely to over- than underrefer, which is appropriate for a screening measure.

Utility: The **ESI** is a quick, easily administered screening instrument that is user-friendly to both children and examiners. The norms are questionable, but the new study should take care of all concerns expressed above. The new version will also include directions for administering the **ESI** to three-year-olds. There are training materials available on videotape. The **ESI** is available in Spanish and Korean.

Availability: Teachers College Press, PO Box 1540, Hagerstown, MD 21740.

Instrument: Florida Kindergarten Screening Battery (FKSB, 1982)

Authors: Paul Satz, Ph.D. and Jack Fletcher, Ph.D.

Purpose: The authors' purpose is to provide a comprehensive screening battery for early identification of children (5-0 to 5-6) with potential learning problems. It is designed to permit mass screening of kindergartners and can be administered by trained paraprofessionals.

Description: The **Florida Kindergarten Screening Battery** is individually administered, requiring about 20 minutes. The **FKSB** is made up of the following five tests:

(1) *Peabody Picture Vocabulary Test* (PPVT; Dunn, 1956), a measure of receptive vocabulary

(2) *Recognition-Discrimination* (Small, 1969), a visual, perceptual (matching to sample) task requiring the child to identify a stimulus geometric design among a group of four figures

(3) *Beery Visual-Motor Integration* (Beery & Buktenica, 1967), an age-normed perceptual-motor copying task

(4) *Alphabet Recitation,* recitation of ABCs, scored by the number of letters named, regardless of order

(5) *Finger Localization* (Benton, 1959), somatosensory test (e.g., recalling the number of fingers touched from the sense of touch alone) consisting of five levels of performance.

The kit for the **FKSP** includes recording forms, a small stimulus book for the recognition-discrimination test and a cardboard screen for the finger localization test. The **PPVT** is not included. If the revised version (**PPVT-R**) is used, the scores should be converted to **PPVT** equivalent scores, using tables in the **PPVT-R** manual.

Scoring: Directions for scoring each of the component tests are included in the manual with the exception of the Beery VMI. Test scores are weighted according to equations derived from the three-year follow-up study. Interpretation of results is discussed in the manual.

Norms: The norms are rated fair. Despite the extensive longitudinal validation information available, the restricted nature of the sample limits the generalizability. They are also dated from 1970.

The **FKSB** was standardized using a longitudinal study that followed 497 children from kindergarten through the elementary school years. The children were all from one county in Florida. Only male children were selected because of the higher incidence of learning problems in that group, and all minority children were excluded because they were likely to be culturally disadvantaged and "representative of the larger population of general academic failure." Approximately 90% of the group came from families in the middle to upper-middle SES levels. Two cross-validation samples were added, one of which did include 28 black children and 20% lower SES. The norms, validity evidence, scoring procedures and weights for the tests must be interpreted relative to the particular characteristics of this sample.

Reliability: The reliability of the **FKSP** is rated fair.

The authors report that the tests that make up the **FKSB** have generally high reliabilities. The authors report most coefficients ranging between .77 and .98. The exception was the finger localization test, on which two subtests had very low reliability, reflecting the influence of ceiling effects.

Validity: Evidence for the validity of the **FKSP** is strong and consistent, however it is rated fair because of limited generalizability.

Content validity: An extensive array of neuropsychological and cognitive tests (13) were administered at the beginning of kindergarten and at the end of grade 2. These were reduced to a smaller subset using multivariate procedures to select the best predictors. That subset (the **PPVT**, a recognition-discrimination test, the **Beery Visual-Motor Integration** test, alphabet recitation and finger localization) comprise the screening battery.

Criterion-related validity: Prediction to academic achievement were based on longitudinal follow-up of the children in the standardization sample. A number of different outcome criteria were used, varying according to the grade level. The results for the three-, six- and seven-year follow-up periods consistently showed good support for the sensitivity and specificity of the battery in predicting severe risk, but relatively poor evidence for predictions of mild risk . The consistent level of predictions of severe risk over the years of the study was impressive, considering the number of years since the original testing.

Teacher ratings were much more accurate in predictions of mild risk, and slightly more accurate in terms of predicting high risk. However this may have been partly due to the low incidence of predictions of high risk. The screening battery had a higher rate of false positives than teachers, but a lower rate of false negatives. That is, the FKSB over- rather than underreferred children, which is desireable in a screening battery.

Construct validity: The first four tests were found to represent separate constructs on the basis of factor analysis. Construct validity of the battery was justified in terms of the range of behaviors tested which were shown to be predictive of poor achievement.

Utility: The **FKSB** is a relatively quick, easily administered screen comprised of five separate tests. The limitations of the normative sample in terms of generalizability may not justify its use over other available instruments. However the extensive research base and longitudinal evidence of predictive validity is rare and commendable.

Availability: Psychological Assessment Resources, Inc. Odessa, Florida 33556.

Instrument: Kindergarten Language Screening Test (KLST, 1983)

Authors: Sharon V. Gauthier, M.A., & Charles L. Madison, Ph.D.

Purpose: The authors' purpose is to provide a screening test encompassing a wide range of tasks reflective of both receptive and expressive language competence. It is based on the verbal language abilities considered normal for children of "kindergarten age." It is designed as a brief, easily administered instrument to discriminate children whose use of language is appropriate for their age and grade from those who have areas of language deficit (as measured by more intensive language testing).

Description: The **KLST** includes 30 items in a format which includes a variety of primarily verbal responses. It is individually administered and requires approximately ten minutes. Specific item content is described as follows:

1. Give full name and age (2 items)
2. Name primary colors (4 items)
3. Count (to 4; to 10) (2 items)
4. Point to body parts (4 items: chin, knee, elbow, ankle)
5. Follow three part sequential command (2 items)
6. Understand prepositions on, under and behind (1 item)
7. Repeat sentences up to 11 words, including conjunctives, interrogatives and embedded clauses (4 items)
8. Spontaneous speech sample noting a variety of speech abilities and syntactic structures (11 items)

Scoring: The total raw score is the sum of items passed correctly. Based on the predictive validity study, the authors suggest that a total score of 19 or below (out of a possible 30) indicates the need for further testing.

Norms: The norms are rated fair because of the limited information provided.

The authors report data for four-year-old norms derived from Headstart children at the time of the test-retest reliability study in 1974. They add this to a larger sample to present "normative data" in six-month intervals (from 48 to 83 months) and provide percentile rankings by age for raw scores in these age intervals. The characteristics of the samples are not described. The mean scores for children under the age of 5 are not based on a sufficient sample to be acceptable as norms.

Reliability: The reliability of the **KLST** is rated fair because of the limited information provided.

Test-retest reliability was .87 in a subsample of 22 children randomly selected from 88 five-year-old Headstart children. Very little information is reported in the manual; the reader is referred to unpublished papers for specifics. Homogeneity of test items was established by a KR-20 reliability coefficient of .86.

Validity: Evidence of the validity of the KLST is rated good.

Content validity: Item selection was literature based and piloted on separate samples of 41 and 113 kindergarten children. The literature on age appropriateness of item content is summarized briefly in the manual. The studies range in date from 1940 (Gesell, et al.) to 1972. The sentence repetition items (including the use of conjunctives, interrogatives and embedded clauses) and the speech sample items are the most thoroughly documented at the kindergarten age range. Low discriminating items were eliminated on the basis of how well each item predicted a child's score on the entire test. Individual item statistics are presented in the manual.

Construct validity: The **KLST** was significantly correlated with the **Utah Test of Language Development** (.60) and three subtests from the **Illinois Test of Psycholinguistic Abilities** (Auditory Reception, .37; Grammatic closure, .36; and Verbal Expression, .40; ITPA sum, .51). The range of scores showed good separation of the upper 25% (mean score of 28/30) and the lower 28% (mean score of 20.4). The SEM was 1.7. The sample for this study included twenty Caucasian and twenty-one Nez Perce Indian children, mean age 6-1. The only significant group difference in this sample was the ITPA grammatic closure test, where Caucasian children demonstrated higher scores. Independent studies reported correlations of .70 between the **KLST** and the **Boehm Test of Basic Concepts**, and of .89 between the **KLST** and the **Clark-Madison Test of Oral Language**.

Predictive validity: The **KLST** was administered to 233 kindergarten children. Thirty of these children received scores below 20 and were tested with the **Northwestern Syntax Screening Test** (mean score below the tenth percentile on receptive and below the third percentile on expressive), as well as the **Boehm Test of Basic Concepts** (mean score at the 29th percentile). Two and a half years later, 82% of the low scoring students were functioning below grade level academically.

Utility: The **KLST** is a quick, easily administered screen for verbal language abilities. With some limitations, there is evidence of construct validity and, more importantly, validity in identifying children who need more in-depth assessment. The use of a variety of language tasks yields a comprehensive picture of expressive and receptive skills and avoids the problem that a single response mode may not match the individual language skills of children.

While the development and early studies appear to include children from a range of SES and ethnic backgrounds, specific information is not provided. It is assumed that this test is appropriate across the range of kindergarten entrants, but it is a significant drawback that the evidence supporting this appropriateness is not presented.

Availability: Pro-Ed, 5341 Industrial Oaks Boulevard, Austin, TX 78735.

Instrument: Miller Assessment for Preschoolers (MAP, 1982, 1988)

Author: Lucy Jane Miller

Purpose: The author's purpose is to provide a screening instrument to identify preschool children in need of more intensive diagnostic evaluation for potential learning problems. The **MAP** is specifically designed to measure differences among children in the lowest 25% performance range and to identify mild, moderate or severe problems that may affect one or more areas of development.

Description: The **MAP** includes 27 "core" items in a format which combines verbal and a variety of performance responses (e.g, block building, drawing, stepping). It is designed for children ages 2 years, 9 months to 5 years, 8 months. The **MAP** is individually administered and requires approximately 25-35 minutes to administer and score. The items are administered with the use of a scoring notebook to hold cue sheets (instructions) and item score sheets, consumable drawing booklets, and a large number of manipulatives which are supplied in the well organized carrying case. Everything needed to administer the **MAP** is provided in the kit, except a stopwatch.

For each item the manual describes the materials needed, the procedure for administration, including standard instructions which should be read exactly as printed (supplemented by the cue sheets), criteria for scoring as well as observations which may supplement scoring. The manual also includes general directions for administration, scoring and interpretation of results as well as cautions about the appropriate and inappropriate use of the test and results. The manual provides a detailed discussion of test development and technical quality.

Care has been taken to make both the administration and the scoring "user friendly" for the administrator as well as the child. Because item administration is different for different age groups, cue sheets and item score sheets are provided for each of the six age groups. The items should be administered in the order presented in the manual. Any change in administration necessitated by the behavior of the child should be noted on the "Behavior During Testing" checklist on the back of the item score sheet. In addition, the behavior checklist allows the examiner to note attention level, social interaction, and sensory reactivity/threshold.

The **MAP** content can be organized into three general ability areas with five Performance Indices (some items fall into more than one Index). The following are examples of specific item content for each area.

Abilities	Performance Index	Number & example items
Sensory & Motor	*Foundations*	10 Items, Sense of position and movement (e.g., hand-to-nose), sense of touch, normal movement patterns
	Coordination	7 Items, Oral motor (e.g., articulation), Fine motor, Gross motor
Cognitive	*Verbal*	4 Items, Cognitive abilities requiring language (e.g., sentence repetition)
	Non-Verbal	5 Items, Cognitive abilities requiring no language (e.g., block designs)
Combined	*Complex Tasks*	4 Items, Visual-Spatial/ Motor Abilities (e.g., draw-a-person, maze)

In addition to screening, the **MAP** provides a comprehensive, structured clinical framework through the use of the Supplemental Observations Sheet. These observations qualitatively describe a child's strengths and weaknesses and indicate possible avenues of remediation. The core 27 items may be adminstered by trained paraprofessionals under the supervision of persons experienced in psychological or developmental assessment. The Supplemental Observations require advanced training.

Scoring: Items are scored according to detailed scoring criteria presented with the administration directions for each item. Item score sheets are customized to each age group and color coded so that the examiner can compare the child's percentile score on each item to that of other children in the same age group. Scores at or below the 5th percentile are coded red ("Stop") and mean the child appears to need further evaluation. Yellow ("Caution", scores between the 6th and 25th percentiles) means that the child should be watched carefully. Green ("Go", scores above the 25th percentile) means that the child seems to be within normal limits .

To obtain the total score, the number of red and yellow scores are recorded. Norm-referenced percentiles for total scores and for performance indices are derived from tables in the manual, based on the number of red and yellow scores on individual items. These percentiles are also categorized by color. The total score categorization (Red, Yellow, Green) can be used as the cutscore for referral. Alternatively it may be more appropriate to chose a different percentile as a cutscore, depending on the specific population and consequences of the cutscore decision. Behavior during testing and supplemental observations can also enter into the decision to refer.

The manual has different scoring criteria for black children which appear to be related to differences in dialect. No rationale is provided for this in the manual.

Norms: Overall the norms are judged to be excellent.

The 1980 standardization sample consisted of 1204 children between the ages of 2-9 and 5-8, approximately 200 children per age interval. The sample was chosen to represent all nine continental geographic census regions of the United States on an approximately equal basis rather than according to population. The sample was judged to be representative of the characteristics of the U.S. population with regard to sex and race. With regard to community size, small towns were slightly overrepresented at the expense of rural areas. The sample is overrepresented by upper socioeconomic (SES) levels based on parental education, job status and family income, with a corresponding underrepresentation of the lowest education and income levels.

The red and yellow cutoff points described above are based on the raw score frequency distribution for each item within each age group. In some cases these were adjusted to better discriminate between the normal and problem population scores.

The final percentile charts for the Total Score and for each Performance Index were obtained by weighting the red and yellow scores for individual items, based on the frequency of these scores in the normative population.

Reliability: Evidence for the reliability of the **MAP** is rated good.

The author presents evidence of three types of reliability. *Stability* of measurement over time (test-retest) was assessed with a sample of 81 children, randomly selected from the standardization sample. The percent of children with the same score category (Red, Green or Yellow) was 81% for the total score, and ranged from 72% to 94% for the performance indices.

Internal consistency reliability was estimated on the basis of raw scores for the total standardization sample at .79 for split-half reliability and .82 for item-to-test correlations. *Inter-rater* reliability was judged from a sample of 40 children who were tested by one administrator and also scored by an observer. The correlations for performance index scores ranged from a low of .84 (due to the articulation item in Coordination) to .97 or above. The fact that the children spanned the entire age range of the **MAP** may have inflated the correlations somewhat.

Validity: Overall, evidence for the validity of the **MAP** is rated good. It would be expected that further research studies with the **MAP** will enhance the evidence of validity.

Content validity: The theoretical foundation and justification of the specific item content of the **MAP** is based on research in a broad range of areas and is well described in the manual. The present content of the **MAP** is a result of 10 years of extensive research involving more than 4,000 children (including children with diagnosed dysfunction) and 800 trial items. 530 items were reduced to the final 27 based on data collected in a nationally sampled item tryout. The tryout sample of 600 normal and 60 preacademic-problem children was stratified on the basis of age, sex, race, size of community and socioeconomic factors. Items were selected on ability to discriminate between age groups, ability to discriminate between normal and children with preacademic problems, to represent a broad range of behavior, and to be easy and inexpensive to administer.

Criterion-related validity: The authors present some evidence of concurrent validity, comparing scores on the **MAP** with performance on the **Wechsler Preschool and Primary Scale of Intelligence (WPPSI)**, the **Illinois Test of Psycholinguistic Abilities** (ITPA), the **Southern California Sensory Integration Tests** and the **Denver Developmental Screening Test** (DDST). The results are somewhat hard to interpret because of significant differences in the purpose and scoring of the criterion tests as well as the small sample sizes. The results do support some level of concurrent validity for the **MAP**. The **MAP** classified more children in at-risk categories than the **DDST**, however the **DDST** is known to underrefer children (Meisels, 1988).

In order to assess predictive validity, approximately one-quarter of the children in the standardization sample were followed up four years after initital testing. Criterion measures of predictive accuracy included a number of standardized tests, report card grades, retention or special placement, and teachers' observations of behavioral problems. The **MAP** total score significantly predicted performance on both intelligence and achievement tests as well as school performance criteria.

In terms of classification analysis, the Red (5th percentile) cutscore had an 8% underreferral rate, with a sensitivity (correct referral) of approximately 20% and a specificity (correct no-referral) of 97%. The Yellow (25th percentile) cutscore had an 5% underreferral rate, with a sensitivity of approximately 51% and a specificity of 79%. While the sensitivity of these cutscores is not particularly high, very few measures predict as well four years from the time of testing. In addition it must be taken into account that there were very few "problem" children in the sample, that the sample covers a broad range of ages both in the initial and outcome testing, and that factors such as school retention policies and early intervention services are not accounted for. The classification accuracy of the **MAP** tended to be better on the criteria which are less influenced by school or district policy.

Construct validity: The construct validity of the MAP was established through factor analysis (matching items to performance indices), the assessment of maturational trends, and item-test correlations. All items are significantly correlated with the total score, and the five performance indices appear to contribute approximately equally to the total score. In addition, 75% of 90 children with established problems were identified by the **MAP** in either the Red or Yellow category. If only children in the upper four age groups are considered (3-9 to 5-8), 84% were correctly identified by the **MAP**.

Utility: The **MAP** is a short, carefully developed, nationally standardized screening instrument. While it may take considerable training for examiners to be familiar with all tasks across all age ranges, the game-like nature of administration should be appealing to children.

The authors demonstrate longitudinal validity for the **MAP** cutscores, although the Red cutscore may have been too conservative (underrefer) for the relatively high SES population studied. Because the normative sample was skewed toward higher parental education levels, the underreferral may be exaggerated. The cutscores may identify a higher proportion of a lower SES population and may prove to be more sensitive with such a population. Examiners working with different populations may want to determine whether a different percentiles would be a more appropriate cutscore for their purposes.

The fact that the **MAP** identifies mild to moderate deviations from normal, rather than just severe problems, may lower the apparent sensitivity. The focus on behavior during testing and supplemental observations, however, should enhance the decision making process for individual children.

A videotape in a programmed learning format is available to ensure that examiners are administering the **MAP** in a standardized manner. Training workshops are available and recommended for use of the Supplemental Observations.

Availability: The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204-2498

Instrument: Pediatric Examination of Educational Readiness (PEER, 1982)

Authors: Melvin D. Levine, M.D., F.A.A.P. and Elizabeth A. Schneider, M.D.

Purpose: The authors' purpose is to provide a multi-dimensional, middle-level screening/diagnostic instrument to identify specific areas of childrens' functioning in need of more intensive diagnositic evaluation for possible learning problems. The PEER functions as a standardized observation procedure in health settings for neurodevelopmental, behavioral and health assessessment.

Description: The Developmental Attainment portion of the **PEER** includes 29 items in a format which combines some verbal and a variety of performance (gross motor, visual/fine motor, neuromotor) responses. It is designed for children ages 4 to 6. The **PEER** is individually administered and requires nearly 60 minutes to administer and score. Some items involve the use of manipulatives (e.g., blocks and cylinders, tennis ball) which are included with the kit; others involve the use of a stimulus booklet which contains instructions and pictures for some of the language and visual-motor tasks. The only materials not included with the kit are a pencil, unlined paper and a penny.

For each task, the manual describes the task, gives instructions for administration (some items have standard oral instructions which should be read exactly as printed), and guidelines for interpretation of the child's response. Responses are recorded on the record form which provides examiner's cues and space for notes after each item. The manual also provides a detailed discussion of interpretations and cautions about overinterpretation and misuse of the results.

Six basic developmental areas are sampled by the **PEER**. The following are examples of specific item content for each area.

Orientation: Identify body parts, imitate finger movements, visual tracking
Gross Motor: Walk on heels, toes, catch ball
Visual-Fine Motor: Matching, copy figures, block construction
Sequential: Finger opposition, object and word span
Linguistic: Spatial directions, complex sentences, categorize
Preacademic Learning: Name days of week, count, write

At three "checkpoints" during the administration of the **PEER**, ratings are made of behaviors under the categories of *Selective Attention/Activity* (activity level, distractibility, fatigability, task persistence, reflective behavior), *Processing Efficiency* (latency of responses, task execution, necessity for instruction and demonstration), and *Adaptation* (rapport with examiner, involvement with examination). The checkpoint system allows the examiner to record differences in these behaviors related to different task requirements (during the tasks requiring primarily motor output, tasks requiring listening and verbal output, and the physical examination).

During the administration of the **PEER** signs that indicate neurological maturation or dsyfunction can be observed and recorded. These signs are discussed in detail for individual items in the manual.

The **PEER** record form facilitates analysis across items of the specific subcomponents of each task with which the child may have difficulty. Adjacent to the scoring section for each item are columns of ten possible task components. The task components considered include four input modes (*visual, verbal, sequential, spatial-somesthetic*), storage (*short-term memory, experience*), and four output modes (*fine motor, motor sequential, verbal sequential, verbal expressive*). The specific components of each task are indicated in columns across the page on the record form by the presence of the numbers (1 2 3) the examiner uses to indicate the level of performance for the item. The performance on specific task components can then be assessed over all tasks by summing the performance levels down the columns.

Scoring: Each task on the Developmental Attainment portion of the **PEER** has three levels of accomplishment indicated on the record form. For each level, the associated scoring criteria are printed on the form. Level One is likely to represent a lag for this age group, Level Two is apt to be appropriate, and Level Three, somewhat advanced. The examiner can also record the child's total inability to approach expectations (Below Levels) or a refusal (Refused Task).

The total number of items at each level is computed and the proportions of Level One, Level Two and Level Three ratings are be determined. A profile is constructed of the results for each area based four levels of concern. These include *Definite Concern* (more than one Level One score), *Possible Concern* (one Level One score), *No Concern* (all items Levels Two and Three), and *Strength* (all items Level 3). Similar levels of concern are recorded for the Associated Observations, determined by the number of appropriate scores.

The neuromaturation findings are rated as *Prominent, Moderate,* or *Few/No Findings* on the basis of numerical scores. Numerical scores indicating *Possible Concern* are also provided for the input, storage and output task analysis results.

The goal of the **PEER** is a narrative description or functional profile rather than a a single overall score, in keeping with the descriptive rather than quantitative nature of children's health assessments. A rating of *Definite Concern* suggests that further evaluation or intervention is required. *Possible Concern* indicates the need for continued monitoring with possible later evaluation or intervention. The **PEER** is not meant to be used in isolation, but should be supplemented with information from parents, teachers or other professionals.

Norms: Overall the normative information is judged to be fair.

The **PEER** is not normed in the traditional sense. "Normative" information provided to determine levels of concern and the levels posted on the record form (One, Two, Three) are based on field testing among predominately middle-class populations. The characteristics and mean levels of performance of one sample are described in two research articles included in the manual. Much of the standardization of the **PEER** has been undertaken with children several months before entry into kindergarten (see sample described below). The scores indicating various levels of concern in the manual and on the record form are based only on children five and older.

Further normative and validation studies were underway in a number of communities at the time the manual was written (1985). The authors state that it is "imperative" that clinicians establish local norms that take into account the nature of educational programs, regional cultural influences, and other conditions that such as language that may have a significant impact on performance.

Reliability: Evidence for the reliability of the **PEER** is rated fair because it is limited.

The authors present evidence of reliability based on the validity study described below. The reliability information was summarized, not presented for every area. The median *inter-rater* reliability was 89%, and agreements ranged from 84% to 95% for separate content areas. No evidence of stability over time (test-retest) reliability is presented.

Validity: Evidence for the validity of the **PEER** is rated good, although the evidence is limited in applicability to middle-class populations of kindergarten entrance age.

Content validity: The **PEER** is more clinically oriented than a developmental inventory or an intelligence test, although the item content may be similar. It was developed under the direction of Dr. Levine at the Division of Ambulatory Pediatrics, The Children's Hospital Medical Center, Boston, MA, and the Brookline Early Educational Project, Brookline, MA.

The content of the **PEER** is based on knowledge of the kind of developmental dysfunctions that can affect children during school years and therefore samples behaviors that are "clinically relevant." This focus should aid in the identification of early predictors of school problems in an efficient manner.

Items are multi-dimensional so that different aspects of functioning can be observed simultaneously. The **PEER** was field-tested with groups of children from two communities, although these samples are not described. Data from the validity study was used to determine that the content of the **PEER** is not redundant.

Criterion-related validity: Evidence of concurrent and predictive validity is presented based on one large sample of children. The sample (part of the **Brookline Early Education Project, BEEP**) consisted of 386 children, 88 of whom were enrolled in **BEEP**. Testing was conducted just prior to kindergarten entry in the summers of 1976 to 1978. The mean age of the children was 61 months (range 53 to 70 months) and approximately 90% had some preschool or daycare experience. Fifty-three percent of the mothers and 65% of the fathers held college degrees. English was the first language in 93% of the homes.

Each child was given the *McCarthy Scales of Children's Abilities*. The mean scores for children with one or more Definite Concern ratings on the **PEER** were significantly lower than those for children in the No Concern group on the General Cognitive Index, as well as all Subtest Indices of the McCarthy.

Kindergarten teachers rated children on mastery, social, academic, gross motor and fine motor skills using the *Kindergarten Performance Profile* (KPP). One hundred eighty-seven children tested on the **PEER** later received fall and spring ratings on the **KPP**. Children with three or more areas rated Definite Concern were rated significantly lower than children in the No Concern group on all areas of the KPP in the fall and spring (with the exception of Gross Motor in the fall). Children with one or more Possible Concern ratings, as well as one or two Definite Concern ratings were significantly lower than children in the No Concern group on the mastery items of the fall KPP (i.e., task persistence, use of time, routines, following directions).

Classification analysis (i.e., sensitivity and specificity) is not presented in the manual. The manual does indicate, however, that while clusters of neurological signs were found to be predictive of later performance, the rates of false-negatives and false-positives were high.

Utility: The **PEER** is a promising instrument which warrants more generalizable validation and standardization. The **PEER** can, and has been used for large scale screening. However, the fact that it takes nearly an hour to administer the entire test, including the physical health and sensory screening portions, make this impractical in many settings. It is more widely used after parents or teachers have expressed concerns about a child. The **PEER** then may be used as part of a diagnostic evaluation or to target specific areas of concern for more intensive evaluation.

The **PEER** was designed to be used by doctors and nurses; however other professional (e.g., psychologists or special educators) can administer all but the physical and neurological components. While the **PEER** is relatively easy to administer and the manual is quite clear, it is important that the examiner have training in children development, familiarity with childhood dysfunctions, and supervised experience in early assessment procedures, interpretations and limitations. The **PEER** has been used as a format for teaching about child development and the low-severity disabilities of childhood.

While referral decisions made on the basis of the **PEER** are not soley based on numerical data, considerable reliance is placed on the "normative" information gathered from five-year-old children. The validity of the concern ratings is partially clinically based; however, forthcoming validity studies with younger children and more generally representative of populations are important to establish confidence in the generalizability of test results.

The fact that the same rating levels are used for children ages 4 through 6 necessitates some difference in interpretation depending on age. A four-year-old would be expected to have more Level One scores than would a child of six. Since many items "ceiling out just before age six," the rating of *Strength* is problematic at that age.

The evidence for reliability and validity is included in the manual in the form of two research publications. This format limits the amount of detail that is provided. It would be helpful to have the reliability and validity information, including classification analysis, presented in more detail in the manual.

Availability: Educators Publishing Service, Inc., 75 Moulton Street, Cambridge, MA 02238-9101

67

APPENDIX C

SUMMARY TABLE OF SCREENING INSTRUMENTS

Content and Key to Instrument Descriptors in Review Summary Tables

INSTRUMENT: *Instrument name, acronym, author(s), publication date and publisher. Indices of instruments by title and publishers' addresses are included after Appendix J.*

FOCUS: *Scope of content covered by the instrument.*

Broad: Includes three or more of the following categories of abilities:
Language, Speech, Cognition, Perception, Personal/Social, Perceptual-motor, Fine, Gross Motor Coordination
Academics: Includes many, but primarily academic skills
Specific Areas: Language, Literacy, Mathematics, Reading, Relational Concepts (see "Content" for specific skills in each area)

AGE/GRADE: *Age or grade range covered by the instrument.*

ADM. TIME: *Time in minutes required for administration and initial scoring.*

FORMAT: *Description of test in terms of type of response required, format and materials, categories are not mutually exclusive*

Format:
- Group or Individual Administration
- Multiple choice
- Paper & Pencil (child marks or writes the answer)
- Stimulus cards/easel
- Manipulatives (e.g., blocks, sorting chips)

Response Mode :
- Teacher rating
- Parent response
- Observation of Child
- Oral (verbal)
- Pointing (implies multiple choice)
- Performance (*fine/visual-motor*: copy, build, write, etc)
- Motor (*gross motor*: hop, skip, jump, catch, etc.)

SCORES: *Types of scores available. No endorsement of the use of specific types of scores is implied* here.

Norm-referenced:
- Percentile, Percentile Rank
- Age Equivalent / Grade Equivalent (Gr.Eq.)
- Standard Score
- Normal Curve Equivalent (NCE)
- Developmental "Age", "Language Age", etc.
- Quotient (Developmental, Language, etc.)

Criterion-referenced:
- Mastery levels
- Raw score

CONTENT: *When the content covers a number of areas, the category name is used. When the content is more limited within a category, the specific areas are named.*

Basic facts:	colors (primary), letters, numbers. shapes
Language:	expressive, receptive vocabulary, fluency, syntax
Literacy:	print functions & conventions, reading symbols
Relational Concepts:	direction, position, size, quantity, order, time, categorization
Listening & Sequencing:	follows directions, remembers story sequences, main ideas
Cognitive:	problem solving, opposite analogies, memory, imitation
Perception:	auditory, visual discrimination
Mathematics:	count rote, with 1/1 correspondence, number skills
Motor:	fine motor (holding a pencil correctly, buttoning, etc) gross motor (hops, skips, throws) visual-motor (copies shapes, builds blocks)
Self:	knowledge of body parts (point or name) social/emotional (peer & teacher interactions, attention span, etc.) self help (buttoning, toilet, etc) information (name, age, address, phone, birthdate)

NORMS: *Ratings on norming studies (value judgement implied)*

None:	no normative information is given
Poor:	some information but limited applicability
Fair:	some standards of comparison (e.g., means of research sample)
Good:	norms based on good sized, representative sample, or lots of relevant information regarding appropriate populations for use
Excellent:	norms based on a representative, national sample and relevant information about applying norms or norm-referenced scores.

RELIABILITY: *Reliability ratings (value judgement implied)*

None:	no reliability information is provided
Poor:	all reliability coefficients (**r**) below .70 or an important type of reliability was not examined
Fair:	at least one reported **r** is greater than .70; or **r** was greater than .80 but evidence was limited in applicability
Good:	total **r** is greater than .80; most subtests have **r** greater than .75
Excellent:	several kinds of reliability reported; total **r** is greater than .90; most subtest scores greater than .80

VALIDITY: *Validity ratings (value judgement implied)*

None:	no validity information is provided
Poor:	information is of very limited applicability
Fair:	most important aspects of were addressed but evidence was moderate or weak; or was strong but limited in applicability
Good:	consistent evidenct of validity, or strong but limited evidence of the type of validity most appropriate for the intended test use
Excellent:	strong evidence and a base of research on the instrument

Summary Table of Instrument Characteristics: Screening Measures

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Basic School Skills Inventory - Screening (BSSI-S) Hamill & Leigh, 1983 PRO-ED	Broad	Ages 4 - 6	5 - 10	Individually Adm Oral & Performance	Basic Facts Counting Speech Fine Motor	Standard Percentile	Poor	Fair Limited	Poor	
Battelle Developmental Inventory - Screening Test (BDI-S) DLM Teaching Resources	Broad	Ages 0 - 8	20 - 30 for ages 3 - 5	Individually Adm Performance Oral, Motor Pointing	Language Cognitive Motor Self	Multiple cutscore probability levels	Poor	None	Fair Limited	Heavily loaded with motor & personal/social items No evidence for technical qualities of cutscores
Bracken Basic Concept Scale - Screening (BBCS-S) Bracken, 1984 The Psychological Corporation	Relational Concepts	Ages 5 - 7	15	Group Adm Paper & Pencil Multiple Choice	Survey of all Relational Concepts	Standard Percentile Stanine NCE	Fair	Fair	Poor Limited	The use of "concept age" score is not recommended
Brigance Preschool Screen Brigance, 1985 Curriculum Associates, Inc.	Broad	Ages 3 & 4	10 - 15	Individually Adm Spiral bound Oral, Pointing Performance	Colors, Motor Language Body Parts Personal data	Raw scores for group ranking	None	None	Content Fair Screening Poor	Parent & Teacher Rating Forms available Not validated for screening
Brigance K & 1 Screen Brigance, 1982 Curriculum Associates, Inc.	Broad	Grades K & 1	10 - 15	Individually Adm Spiral bound Oral, Pointing Performance	Basic Facts Language Mathematics Motor	Raw scores for group ranking	None	None	Good Limited	Parent & Teacher Rating Forms available Author has not validated this test for screening
The Communication Screen Striffler & Willig, 1981 (TCS) Communication Skill Builders	Language	Ages 2,10 to 5,9	2 - 5	Individually Adm Stimulus card Oral & Perform. Observations	Language Cognitive	Pass Suspect Fail	Preliminary Limited	Fair Limited	Fair Limited	Developed by clinicians Needs more evidence of technical quality, smaller age groups for scoring
Denver Developmental Screening Test (DDST) Frankenburg et al., 1975 LA-DOCA Project & Publishing Fndtn	Broad	Ages 0 - 6	20	Individually Adm Manipulatives Motor, Oral Performance	Self Fine Motor Language Gross Motor	Cutscores	Poor Dated	Fair Limited	Fair	Conservative test, errs on the side of underreferrals

Summary Table of Instrument Characteristics: Screening Measures cont.

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Developmental Activities Screening Inventory II Fewell & Langley, 1984 (DASI II) PRO-ED	Primarily Academics	Ages 0 - 5	Untimed	Individually Adm Pointing Performance few Oral	Colors Classify Visual Motor Memory Spatial Reltns	Developm. Age & Quotient	None	None	Poor	
Developmental Indicators for the Assessment of Learning-Revised (DIAL-R) Childcraft Education Corporation	Broad	Ages 4 - 6	5 - 10	Individually Adm Oral & Performance	Basic Facts Counting Speech Fine Motor	Standard Percentile	Fair	Fair Limited	Fair	
Early Identification Screening Program (EISP) Baltimore City Public Schools, 1982 Modern Curriculum Press	Academics	Grades K & 1	20	Individually Adm Performance Oral	Perception Colors (name) Shapes Visual Motor	Total raw score	None	Good	Fair	
Early Screening Inventory (ESI) Meisels & Wiske, 1983 Teachers College Press	Broad	Ages 4 - 6	15 - 20	Individually Adm Performance Motor & Oral	Cognitive Counting Language Motor	Cutscores: OK Rescreen Refer	Fair	Good Limited	Good	Extensive new norm study underway includes 3-year-olds
Florida Kindergarten Screening Battery (FKSB) Satz & Fletcher, 1982 Psychological Assessmt Resources	Language Perception	Grade K	20	Individually Adm Oral Performance	Vocabulary Visual Motor Perception Alphabet	Individual test scores are weighted	Fair	Fair	Fair	Impressive longitudinal validity studies but of limited generalizability
Fluharty Preschool Speech and Language Screening Test Fluharty, 1978 DLM Teaching Resources	Language	Ages 2 - 6	6	Individually Adm Picture cards Oral Pointing	Vocabulary Articulation Comprehension Repetition	Cutscores for each subtest	Good	Good Limited	Unclear	Specific instructions on how to make allowances for Black dialect Cutscore develop. unclear
Kindergarten Language Screening Test (KLST) Gauthier & Madison, 1983. PRO-ED	Language	Grade K	10	Individually Adm Oral	Basic Facts Language Self Follow Direction	Total Raw score	Fair Limited	Fair Limited	Good	Measures a broad variety of language skills

Summary Table of Instrument Characteristics: Screening Measures cont.

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
McCarthy Screening Test (MST) McCarthy, 1978 The Psychological Corporation	Broad	Ages 4 - 6 1/2	20	Individually Adm Manipulatives Motor, Oral Performance	Motor Cognitive Language Mathematics	Pass/Fail by subtest Cutscores: # failed	Good Dated	Fair Limited	Good Limited	Developed from MSCA No independent norms validity or reliability
Miller Assessment for Preschoolers (MAP) Miller, 1984 The The Psychological Corporation	Broad	Ages 2,9 to 5,8	25 - 35	Individually Adm Motor Performance Oral	Broad range of Motor and Language Skills	Percentile cutscores	Excellent	Good	Good	Training video available Supplemental behavior observations
Mullen Scales of Early Learning (MSEL) Mullen, 1984 T.O.T.A.L. Child, Inc.	Broad	Ages 1,3 to 5,8	35 - 45	Individually Adm Manipulatives Picture Books Oral & Perform.	Perception Language Cognitive Visual Motor	Age scores T-scores	Good	Good	Good Limited	Test materials include colorful toys attractive to children
Pediatric Examination of Educational Readiness (PEER) Levine & Schneider, 1982 Educators Publishing Service	Broad	Ages 4 - 6	60	Individually Adm Performance Oral, Motor	Language Basic Facts Motor Orientation	Concern Level cutscores	Fair	Fair Limited	Good Limited	Designed for medical setting or interdisciplinary screening
Preschool Development Inventory (PDI) Ireton, 1984 Behavior Science Systems	Primarily Academics	Ages 3 - 5 1/2	25	Individually Adm Parental rating Yes/No format	Language Motor Self, Social Problem behav	cutscores	Fair Limited	None	Poor Limited	
Screening for Related Early Educational Needs (SCREEN) Hresko et al., 1988 PRO-ED	Academics	Ages 3 - 7	15 - 40	Individually Adm Pointing, Oral Performance	Language Reading Writing Mathematics	Standard Percentile	Good	after age 6 Good Limited	Fair	Little evidence of reliability and validity is poor for the 3-5 age range
SEARCH Silver & Hagin, (1981) Walker Educational Book Corporation	Perception	Ages 5,3 to 6,8	20	Individually Adm Manipulatives Performance Oral, Motor	Perception Perceptual/ Motor, Memory Articulation	Ability Profile Stanines Cutscores	Fair Dated (1973)	Fair Limited	Fair Limited	Multiethnic content depiction

APPENDIX D

SELECTION CHECKLIST FOR READINESS MASTERY INSTRUMENTS

Selection Checklist for Instruments Measuring Mastery of Readiness Skills

I. Utility

A. Information Obtained

1. Is the stated use of this instrument to provide information on mastery of specific skills in a manner that it can be used for individual and/or group planning?
2. Does the instrument provide results or scores which are easily calculated, readily interpreted, and *useful for instructional planning* in terms of the specific curriculum content?
3. Does the manual provide information on the interpretation of results in terms of instructional planning?
4. Does the instrument appropriately cover the entire range of skill that can be expected (i.e., no "ceiling" or "floors" in terms of scores)?
4. Does the instrument provide help with reporting to parents and/or other educational professionals?
5. Is the instrument available and validated for the languages needed in your community?

B. Logistics

1. Can the instrument be administered in a reasonable length of time, considering the amount and quality of information it provides?
2. Is the instrument easy to use? Who can administer the test (teachers, specialists, trained assistants), and what kind of training will be necessary?
3. Are training materials provided?
4. Are any special facilities and/or equipment needed for administration?

C. Cost

1. Are the costs within available resources? Include costs of obtaining the instrument (manual, test kit, consumable test forms, record sheets, etc.), training administrators, and the time to collect and record results.

II. Validity

A. Evidence for Content Validity

1. Is the content appropriate to provide information that will be useful for the specific planning or documentation task at hand? How was the content determined in the test development process? Has the content been reviewed by experts?

2. Does the content completely cover what you intend to measure, or are there important areas not covered?

The following list is a guide for general academic readiness skills based on the *World Book* (1987) survey of more than 3000 kindergarten teachers throughout the United States and Canada on the skills and knowledge a child needs in order to begin kindergarten successfully.

Knowledge of basic facts: colors, letters, numbers, shapes
Language: expressive, receptive vocabulary
Emergent literacy: concepts of word, sentence,
communication structures
Relational concepts: classification, categorization
mathematical (more, less, first, second, etc.)
position (on, under, etc.)
size (big, long, etc.)
Counting: one-to-one correspondence, rote
Listening & Remembering: follows simple directions
remembers story sequences and ideas
Personal, social/emotional: point to body parts
social/emotional (taking turns, sharing)
self-help, able to give own name(s), age

An important consideration is whether the instrument provides for parental input.

3. Does the child understand what she/he is being asked? Is there evidence that the instructions, the format, and the response required are appropriate to measure what is intended rather than attention span, cultural background, ability to speak English, etc.?
4. Will the assessment experience be pleasant for young children?

B. Evidence for Criterion-Related Validity

Is there evidence that this measure is related to other similar and valid measures?

III. Reliability

A. Is there evidence of *stability* over time (*test-retest*)?
B. Is there evidence of *inter-rater* reliability?
C. Is there evidence of *internal consistency* within the test or within subtests?

IV. Norms

A. Is the test norm-referenced? Are the derived scores related to home or preschool experience rather than age?
B. Was the size of the norm group sufficient to have confidence in the norms? In particular, were there reasonable numbers of children in each age group?
C. How similar are the characteristics of the norm group (e.g., sex, race, geographic location, parental education) to the population which will be screened?

APPENDIX E

REVIEWS OF READINESS MASTERY INSTRUMENTS

Contents of Appendix E

Page			
1		Analysis of Readiness Skills	
4	**BSSI-D**	Basic School Skills Inventory	Diagnostic
7	**Boehm-R**	Boehm Test of Basic Concepts-Revised	
11	**Boehm-PV**	Boehm Test of Basic Concepts-Preschool Version	
15	**BBCS-D**	Bracken Basic Concept Scale, Diagnostic	
19	**CSAB**	Cognitive Skills Assessment Battery	
21		The Lollipop Test	
24	**SRS**	School Readiness Survey	
26	**TELD**	Test of Early Language Development	
28	**TEMA**	Test of Early Mathematics Ability	
30	**TERA**	Test of Early Reading Ability	
32	**TOLD-2**	Test of Language Development, Primary	

Instrument: Analysis of Readiness Skills (1972)

Authors: Mary C. Rodrigues, William H. Vogler and James F. Wilson

Purpose: The authors' purpose is to provide an instrument to assess the mastery of basic concepts in reading and mathematics as an aid for instructional grouping and planning. The test content is limited to matching and identification of letters, identification of numbers, and counting.

Description: *The* includes 40 items in a multiple choice format which requires the child to mark the correct response on a test booklet. It is designed for children entering kindergarten, between the ages of 5 and 5 1/2. It can be administered individually or to groups of up to 15 students and requires approximately 30 to 40 minutes to administer. The items are administered with the use of an eight-page consumable test booklet. Each of three subtests consists of two pages in the record booklet, with five items per page. Letters and numbers are printed in large black type; items are outlined with green frames. A marker strip may be necessary to help children keep their place in the booklet. Responses are recorded by the child in the test booklet by marking the response choice with an X. A sample item chart is provided to demonstrate sample items. A class record sheet is also provided in the manual.

The manual contains standard instructions for administration, in both English and Spanish, which should be read exactly as printed. The manual includes general directions and precautions for administration, scoring and interpretation of results as well as a brief discussion of test development and technical characteristics.

The **Analysis of Readiness Skills** consists of three separate subtests. The following are descriptions of specific item content for each area.

Visual Perception of Letters: Match one of five uppercase or lowercase letters to a separate stimulus letter

Letter Identification: Mark one of five uppercase or lowercase letters named by the teacher

Mathematics: Each item consists of four numbers and three sets of black dots and is scored for two activities: (a) mark the number named by the teacher and (b) mark the set of dots corresponding to that number

Scoring: Items are scored on a pass/fail basis according to the answer key provided in the manual. Total raw scores can be converted into percentile ranks, which are presented separately for the English and Spanish normative samples. High, medium and low score ranges are presented for each subtest and for the total test for the English sample. These represent the upper 37.5%, the middle 37.5% and the bottom 25% of the normative group. According to the authors, children scoring in the high range can immediately benefit from reading and mathematics programs. Children scoring in the medium range should have a six- to eight-week interval of in-depth readiness experiences and activities before starting structured programs. Children scoring in the low range should not be placed in structured programs until subsequent testing and teacher judgment confirm they are ready.

Norms: Overall, the norms are judged to be poor, largely due to the fact that they are nearly twenty years out of date (September - October 1971).

The English-speaking standardization sample consisted of 3,305 children beginning kindergarten, from 17 states. Approximately 10% of the children in the schools sampled did not complete the test or were not tested because teachers did not feel they were ready for testing. Thus, the authors suggest, it is likely that the sample does not represent the lower 10% of the kindergarten population.

The sample was judged to be representative of the characteristics of the U.S. population with regard to geographic region and community size, with the exception of an underrepresentation of the Southwest. While socioeconomic status (median education and family income within community) was considered in the sample plan, no specific information is provided in the manual. No information is provided as to the actual age range of the children.

The Spanish-speaking standardization sample consisted of 685 children beginning kindergarten, from seven states. These children attended school districts in which at least 5% of the population were identified as Spanish-American. Approximately 20% of the children in the schools sampled did not complete the test or were not tested because teachers did not feel they were ready for testing. Thus it is likely that the sample does not represent the lower 20% of the population.

It is not specifically stated whether the normative sample was tested in groups or individually. This has relevance for the application of the norms.

Reliability: Evidence for the reliability of the **Analysis of Readiness Skills** is rated poor, largely because stability of measurement over time (test-retest) was not assessed.

The authors present evidence only of internal consistency reliability, which was good for the English-speaking sample and fair for the Spanish speaking sample. For the English-speaking sample the internal consistency was .90 for the total test, and ranged from .59 to .87 for separate subtests (the lowest value probably due to restricted range of scores in the letter matching subtest). For the Spanish speaking sample the internal consistency was .81 for the total test, and ranged from .54 to .71 for separate subtests.

Validity: Virtually no evidence is presented to support the validity of the **Analysis of Readiness Skills** and therefore it is rated poor.

Content validity: No justification for item content is given beyond the statement that research studies indicate knowledge of the alphabet and numbers are reliable indices of reading and mathematics "readiness." Two relatively large pilot studies were done to provide information for item analysis.

Criterion-related validity: The predictive validity of a pilot form of the **Analysis of Readiness Skills** was assessed in relation to teacher judgements and scores on the *Metropolitan Readiness Test* at the end of kindergarten. The authors report that the two tests correlated highly with each other, but only moderately with teacher judgments. However, no specific details (correlations, numbers of children) are given. There is no indication of what the differences were between the pilot and the final version of the **Analysis of Readiness Skills**.

Utility: The **Analysis of Readiness Skills** is a quick, easily administered test of letter and number knowledge which has limited applicability for instructional planning. The normative information is seriously outdated and there is no evidence for the validity of instructional placement decisions based on the high, medium, and low score ranges.

The **Analysis of Readiness Skills** covers a narrow range of skills, which indeed may have been necessary prerequisites to the type of structured reading and mathematics instructional programs prevalent at the time the test was developed. It could provide an objective format to assess letter and number knowledge within a kindergarten class; however, there are more recently normed instruments which also provide this information. The paper and pencil test format, with multiple items on a page, is not appropriate for preschool children.

Availability: The Riverside Publishing Company, 8420 Bryn Mawr Avenue, Chicago, IL 60631.

Instrument: Basic School Skills Inventory - Diagnostic (BSSI-D, 1983)

Authors: Donald D. Hammill and James E. Leigh

Purpose: The authors intend this as a dual-purpose instrument. As a norm-referenced measure of early abilities related to daily living skills, spoken language, reading, writing, mathematics and classroom behavior, it is intended to be used to identify children in need of comprehensive diagnostic evaluations. The information can also be used in a criterion-referenced assessment for instructional planning and monitoring of progress.

Description: The **BSSI-D** includes 110 items in a format which combines oral and performance reponses from children with teacher ratings. It is designed to be used with children ages 4 to 6. The **BSSI-D** is individually administered and requires approximately 20-30 minutes, depending on how well the administrator knows the child and the test. The manual contains standard instructions for administration, directions for scoring and interpretation and the normative data tables. Responses are recorded on the Pupil Record Form, which also provides a chart for creating a profile of the standard scores for each subtest. A picture book is used for direct testing on the spoken language, reading, and mathematics subtests.

The **BSSI-D** covers the following areas in six separate subtests:

Daily Living Skills: primary self-care behaviors (e.g., washing, buttoning)
motor behaviors related to school activities (cutting, folding, drawing shapes)
independent and responsible behavior
basic information (telling time, days of the week)

Spoken language: appropriateness of vocabulary, use and structure of language

Reading: letter knowledge, sound-symbol relationships, predict words from context, early literacy skills

Writing: writing letters, copying words and sentences, spelling, capitalization and punctuation, composing

Mathematics: recognition and printing numerals, counting, quantitative relationships, equivalence, seriation, simple computation

Classroom behavior: attentiveness, cooperation, attitude, socialization, work habits

Scoring: Items are scored on a pass/fail basis, according to scoring criteria presented in the manual for each item with the instructions for administration. For some items, the administration directions are standard and the scoring criteria are objective. For others, the teacher scores the item on the basis of knowledge or observations. In many items of this type, the scoring criteria are extremely subjective. For example, in assessing whether vocabulary is age appropriate, one of the criteria for not giving credit is that the child "*seem[s]* to have a restricted or `immature' vocabulary in comparison with other chldren in the class." Items that require the teacher's intepretation of terms like "appropriate" are particularly problematic on a norm-referenced test.

The child's raw score can be converted into standard scores and percentiles for each subtest and for the total, using tables in the manual.

Norms: Overall, the norms are judged to be fair.

The standardization sample consisted of 813 children between the ages of 4-0 and 7-5 from 18 states. The sample was judged to be representative of the characteristics of the U.S. population with regard to sex, race, and urban/rural residence. In terms of parent occupation, blue-collar workers were over represented (66% sample compared to 36% population), with a corresponding underrepresentation of white-collar workers. In terms of regional distribution, the West is seriously underrepresented (2% sample compared to 19% population), with a corresponding overrepresentation of the South.

The derived standard and percentile scores are based on the average scores of the standardization sample in each six-month age interval from 5-0 to 7-5; the 4-0 and 4-6 age intervals were combined. No information is presented as to the numbers of children tested at each age range and the mean and standard deviation of scores for the total sample.

As was mentioned, the lack of standard administration procedures and objective scoring criteria for many items makes norm-referenced interpretation questionable.

Reliability: Overall, the reliability is rated fair because of the lack of evidence for these important types of reliability.

The author(s) present evidence of high *internal consistency* reliability, ranging from .79 to .97 across ages and subtests. *Alternate forms* reliability was examined on standardization scores between the **BSSI-D** and the shorter screening form, the **BSSI-S**. The correlations were .91, .92 and .88 for ages 4, 5 and 6, respectively. However, if both forms were scored by the same teacher at the same point in time, these correlations may represent a substantial overestimate of reliability.

Stability of measurement over time was not examined, nor was *inter-rater* reliability. Because, as the manual states, the **BSSI-D** is "to an extent a measure of teacher's perceptions of children's abilities and specific skills" (p. 15), the lack of evidence of inter-rater reliability is a serious issue.

Validity: Evidence for the validity of the **BSSI-D** is rated poor.

Content validity: The **BSSI-D** is a revision of the 1976 *Basic School Skills Inventory* (**BSSI**). Although the manual states that the 1983 edition was altered considerably and field tested twice, the numbers and characteristics of the pilot samples are not described. Only teacher opinion is offered to justify the specific content.

The original **BSSI** was based on opinions of 50 kindergarten and first grade teachers on what the distinguishing educational and behavioral characteristics were for actual children they considered "ready" and "unready." This 67-item form was field tested twice and revised on the basis of item analysis, reliabliity data, and teachers' suggestions, then nationally normed. The items were assigned to subtests "on the basis of face validity." Criteria for item selection included that the skill be directly related to school performance, teachable, and not directly related to the home environment or health of the child.

Item difficulty and discrimination statistics were used to select items. The mean item difficulty and discrimination statistics are presented in the manual for a random sample of 120 children in the standardization sample. These statistics indicate that, except for four-year-old writing performance, the items are appropriate for the ages of 4-6. There appears to be a ceiling effect on most subtests after age 6-0.

The problems with subjective administration and scoring criteria affect the content validity in terms of the appropriateness of the manner in which the content is measured.

Criterion-related validity: Concurrent validity of the **BSSI-D** was evaluated in relation to teacher ratings. The correlations of teacher ratings with **BSSI-D** subtests, while statistically significant, were small (.22 to .38; .43 for the total test). The value of this evidence is questionable since the ratings were for "general readiness" on a three-point scale, and the scores of the test were also largely teacher perceptions.

Construct validity: General evidence is presented supporting the relationship between **BSSI-D** and chronological age, as well as the relationships among subtests. The authors interpret this as an indication that the **BSSI-D** measures "readiness" as a developmental construct consistently across subtests. Evidence that the **BSSI-D** differentiates children diagnosed as "learning disabled" from "normal" children was presented for a sample of 12 children.

Overall, the validity for the **BSSI-D** as a measure to identify children with potential learning problems is rated poor because of the lack of information on the sensitivity and specificity of classifications based on **BSSI-D** results. As a measure of readiness, the validity judgement depends on the specific application; however, the evidence of content validity is limited. Teachers' characterizations of the behaviors displayed by "unready" children should not be interpreted as evidence of a cause-effect relationship. There is no evidence, for example, that the ability to cut with scissors has any relationship with school success.

Utility: There are several problems with the **BSSI-D** that seriously limit its utility for either purpose, but particulary as a norm-referenced screening device. The most important concern that limits the utility is the questionable nature of the norms. Ceiling effects, particularly at the upper age ranges, limit the interpretability of scores, because no child can perform above "average" on a number of subtests. The lack of evidence of inter-rater reliability, as well as criterion-related validity, contributes a lack of confidence in the results, whether they are interpreted as norm- or criterion-referenced.

The lack of justification for specific content, and particularly for the combinations of skills into subtests, limits the interpretability of the scores even as a readiness measure. What does a number mean that represents a combination of skills such as tying shoes, folding paper and naming the days of the week?

Because the **BSSI-D** does not differentiate skills that represent acquired knowledge from the underlying ability to aquire knowledge, it furthers the confusion regarding the functions and separate focus of screening and readiness tests.

The **BSSI-D** would require considerable training to administer smoothly, due to the way in which items are presented, alternating between different response requirements. The **BSSI-D** must be administered by someone familiar with the child's classroom behavior.

Availability: Pro-Ed, 5341 Industrial Oaks Blvd., Austin, Texas, 78735.

Instrument: Boehm Test of Basic Concepts, Revised (Boehm-R, 1986)

Author: Ann E. Boehm

Purpose: The author's purpose is to provide a measure of children's mastery of the language of instruction. That is, "those concepts considered basic to understanding directions and other oral communications at the preschool and primary grade level, and to using materials that are designed to teach reading and basic mathematics at these levels" (p. 59). The results of the **Boehm-R** can be used to plan instruction for an individual child whose overall level of concept mastery is low, or to plan group instruction for individual concepts with which large numbers of children within a classroom may be unfamiliar.

Description: The **Boehm-R** consists of 50 items presented in a multiple choice format. The teacher reads aloud a statement that is true of one picture (e.g., "Mark the tree with the bird at the *bottom*") and the children mark (in pencil or crayon with a large X) the one correct picture out of three alternatives. Standard instructions for administration should be read exactly as printed in the manual. The **Boehm-R** is a group administered test. It is recommended that young children be tested in small groups and/or teachers' aides be used to assist children.

The items are arranged in approximate order of increasing difficulty and divided evenly into two 25-item booklets. Each booklet takes 15-20 minutes to administer to kindergarten classes, including the time needed for general instructions and three sample items. The booklets can be administered in separate sessions. Two alternate forms are available, Forms C and D, aiding in pre- and post-testing situations.

The basic concepts addressed by the **Boehm-R** are the relational concepts having to do with *space* (location, direction, orientation, dimensions), *quantity* (and number), and *time*. These concepts have been identified as those needed by children to:

- "understand and describe relationships between and among objects, the locations and characteristics of persons, places and things, and the order of events [e.g., different, between, last];
- follow teacher directions [e.g., top, left]
- comply with the demands of instruction in the areas of language arts, mathematics, and science [e.g., more, first];
- comply with the procedural aspects of teacher-made and standardized tests [e.g., beginning, skip]; and
- engage in problem-solving activities that involve classifying, sequencing, comparing, and identifying multiple attributes [e.g., every, second, alike, as many]." (p. 2)

An optional applications booklet is available for use with the **Boehm-R** in grades 1 and 2. This 26 item booklet, based on actual transcripts of teachers' directions to students, addresses children's ability to use the concepts included in the **Boehm-R** in such tasks as the following:

- following multiple step directions
- making comparisons to a standard
- making comparisons involving an intermediate position
- placing objects and events in order

The manual presents administration and scoring procedures, and suggestions for interpretation of results and instructional planning, as well as tables for scoring and technical information about the test development.

Scoring: The class record form serves as a scoring key as well as guide for interpretation of results. A scoring matrix with miniature reproductions of each item down the page (marked to show the correct response), and places for children's names across the top of the page, allows the teacher to accumulate information for the entire class while scoring each test booklet. The child's score is the total number of items answered correctly (added down the column).

The teacher can examine the level of mastery for specific concepts by counting the number of children in the class answering each item correctly across the columns. Analysis of the class average and the percent of students passing each item can be used to determine whether group or individual instruction is needed on specific concepts.

Children's raw scores can be converted into percentiles (and NCEs) for each grade and time of year, using the national norms presented in the manual. Optionally, an analysis of the type of children's errors may be made (e.g., no response, marking an antonym of the target concept, marking every picture).

Norms: The norms are rated excellent because of the size and representativeness of the sample, and the variety of information provided. The presentation of item data is particularly useful.

The standardization sample consisted of approximately 5,000 children, in kindergarten, grades 1 and 2, for each form and for each time of testing (fall, spring). Forms C and D were administered to children in regular public school classrooms. The sample design was based on national statistics published by the Center for Educational Statistics. Scores were statistically adjusted to make the overall sample match national school enrollment data for district size and region.

The socioeconomic level (SES) of the sample participants was estimated on the school level, rather than recorded for individual students. The percentage of children who qualified for subsidized lunches was used as a proxy for family income in assigning the SES levels which were used in subsequent analyses. The mean SES Indices for the sample were comparable to those provided in the 1980 U.S. Census.

The percentage of children passing each individual item (item difficulties) are presented for fall and spring testing, for kindergarten, grade 1 and grade 2. These item difficulties are presented for the total group and separately for the lowest SES level (50% or more children eligible for subsidized lunches), and combining the mid and highest SES levels. Raw scores can be converted into percentiles for each grade.

The **Boehm-R** has appropriate levels of difficulty and shows evidence of being an appropriate measure of growth in conceptual skills for kindergarten children. There appears to be a significant ceiling effect for grades 1 and 2, with consequences for the usefulness of the **Boehm-R** at those grade levels, as well as the reliability data presented below. The table which follows presents the mean raw score, in terms of the percentage of items passed, and for the total group at each grade level as well as the low SES group.

	Kindergarten		Grade 1		Grade 2	
	Low SES	Total	Low SES	Total	Low SES	Total
Mean Percent Items Passed:						
Fall (beg. of year)	68	75	86	90	93	95
Spring (end of year)	82	85	88	92	94	96

Examination of the range of item difficulties presented in the manual for each item, as well as the means and standard deviations, indicates that many of the students beginning grade 1 and almost all of the students beginning grade 2 can answer nearly every item on the test correctly. As a mastery test, the **Boehm-R** will identify a small minority of low scoring children at these grade levels, but it may not be useful for the majority of children. It would not be an appropriate measure for measuring growth in grades 1 and 2, since there is little variation from fall to spring. The applications form, while too difficult for kindergarten, is appropriate for grades 1 and 2 in terms of item difficulty.

Reliability: Overall, the evidence of reliability is judged to be fair. However, since the relatively low correlations were probably due to the ceiling effect, for kindergarten specifically the reliability is judged to be good.

The reliability of scores between the *alternate forms* was established with correlations of .82, .77 and .65 for kindergarten, grade 1 and grade 2, respectively. The correlations for grades 1 and 2 are not very good, but very probably they decrease from kindergarten to grade 2 because of the ceiling effect. That is, when the scores are clustered at the top of the scale (lack of variability within each form), it limits the size of the correlation that is possible. *Internal consistency* was measured by split-half reliablity coefficients ranging from .85 to .64 (grade 2). Again, lower correlations were probably due to the ceiling effect.

Test-retest reliability was examined in a separate study of approximately 200 children per grade level. These children were administered the same form of the **Boehm-R** twice, approximately one week apart. Reliability coefficients ranged from .88 (kindergarten) to .75 (grade 2), with one outlier of .55 (grade 1, Form D).

Validity: The validity of the **Boehm-R** is rated excellent in terms of its use as a measure of concepts related to the language of instruction *in kindergarten*, and good overall. This rating is based on the extensive focus on content validity (including pilot testing) and the moderate predictive relationship with achievement.

Content validity. Content validity, i.e., the representativeness of the items in terms of the basic concepts essential for school success, was considered the most important type of validity for a mastery test of this type. The author spent considerable effort to acquire and to present evidence of content validity in terms of a universal set of basic concepts gleaned from many sources (briefly summarized below). In terms of use as a mastery test, a pre- post-test of instructional effectiveness or as a measure of readiness for standard classroom instruction, the match between the set of concepts addressed in the **Boehm-R** and those taught or required in the specific program using the test is the ultimate measure of content validity.

In this revision of the *Boehm Test of Basic Concepts* (BTBC), the importance of each original concept was reassessed in relation to the frequency of occurrence in printed materials, reading and mathematics curricula, teachers' verbal instructions and comments by users of the BTBC. More than 1,500 children from a wide geographic distribution of states participated in field testing in which new and revised items were tried out for difficulty, clarity of concept and relationship with the BTBC. Six items from each of Forms C and D were dropped after the standardization on the basis of item analysis results, teacher comments, or reviews by members of a bias panel.

Criterion-related validity. One study is presented as evidence of predictive validity. Three school districts involved in the spring standardization provided individual scores on standardized achievement tests one year after the **Boehm-R** testing. The correlations between **Boehm-R** and achievement scores ranged from .28 to .64 with a median of .44. The strongest correlations were for kindergarten children. The means for grades 1 and 2 indicated that a ceiling effect on the **Boehm-R** restricted the correlations with achievement in these grades.

Other evidence of validity comes from the research base available for the BTBC, offering evidence for predictive validity in terms of achievement, readiness and language, and evidence for construct validity in terms of the sensitivity of the concepts to instruction. Evidence for the validity of the **Boehm-R** in grades 1 and 2 would be enhanced by studies including the applications form, which is more appropriate in terms of item difficulty.

Utility: The **Boehm-R** is a relatively quick, easily administered test of basic concepts particularly appropriate for use with kindergarten children. It has outstanding technical quality at the kindergarten grade level for use as a mastery or specific "readiness" type of assessment. It has more limited utility for grades 1 and 2 due to a significant ceiling effect.

The **Boehm-R** has been used for pre-kindergarten children in an individually administered instrument. While there is some information available about its use for this age group, the authors recommend use of the preschool version which has been standardized with three- and four-year-old children. The original form of the **Boehm-R**, the BTBC, has been used effectively for children with a variety of physical and cognitive handicapping conditions. A number of studies examining sex, SES and cultural bias indicate that, as a whole, the content of the BTBC is not biased toward particular groups.

The class record form is well designed for aid in scoring and summarizing class results. A Parent-Teacher Conference Report form is available which includes a brief description of the test and other information to aid in explaining the test results to parents.

An excellent discussion of the importance and selection of concepts as well as suggestions for interpreting performance and strategies for instruction are presented in the manual. Instructions and cautions about pre- and post-testing are also clearly presented. While the **Boehm-R** measures children's ability to respond to concepts in a print format, no difference was found between scores on a printed versus an object version with Head Start children (Ault, Cromer & Mitchell, 1977). There is a Spanish version available.

Availability: The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204

Instrument: Boehm Test of Basic Concepts, Preschool Version (Boehm-PV, 1986)

Author: Ann E. Boehm

Purpose: The author's purpose is to provide a measure of young children's mastery of basic concepts as an indicator of school readiness and as a guide for planning language instruction.

Description: The **Boehm-PV** consists of 52 items presented in an individually administered, multiple choice format which requires the child to point to the correct answer. The teacher reads aloud a statement that is true of one picture (e.g., "Point to the cat *on* the box.") and the child points to the one correct picture out of several (usually three) alternatives. The pictures for the items are presented in a spiral-bound picture book. The picture book forms an easel between the child and examiner, with pictures facing the child and the standard instructions for the item facing the examiner. The standard instructions for administration should be read exactly as printed in the picture book.

The **Boehm-PV** is designed for children three to five years of age and takes about 15 minutes to administer. Twenty-six basic relational concepts are addressed (two items per concept), having to do with *size, direction*, *position in space*, *quantity*, and *time*. The manual stresses that all 52 items must be administered so that the child has two chances to demonstrate understanding of each concept. The items are arranged in approximate order of increasing difficulty. There are five warm-up items (A - E) and testing is discontinued if the child is not able to answer two of warm-up items B through E.

The manual presents administration and scoring procedures, and suggestions for interpretation of results and instructional planning, as well as tables for scoring and technical information about test development.

Scoring: Each item is scored on a pass/fail basis. Scores for each of the 26 concepts are calculated by summing the responses for the two items tapping that concept, and therefore range from 0 to 2. The child's responses are recorded on an individual record form. The illustrations for each item in the picture book have been reproduced on the individual record form and ordered in such a way that the two items for each concept are side by side. This arrangement facilitates entering the "concept score" as the sum of the two item responses. The 26 concept scores are summed to yield a total score.

Children's total scores can be converted into percentiles (and T-scores) for each age interval, using the national norms presented in the manual. Optionally, an analysis of the consistency of antonym response selection (i.e., confusing the concept with the opposite) may be made. A special section is included on the individual record form to aid in this process.

A class record form is available for planning group instruction.

Norms: The norms are rated fair.

The standardization sample included 433 children, averaging 86 children in each of five age intervals. Children enrolled at 35 sites in 17 states were tested, beginning in early 1985 and ending in the spring of 1986. The 35 sites included private day-care centers, nursery schools, public preschools and Head Start programs. No details are given as to the distribution of children from different types of programs, nor how long the children had been enrolled in a given program. The sample was selected to be representative of the U.S. population in terms of race, geographic region and educational level of parents. With the exception of some divergence for the three-year-olds (less than ten percent), the sample was well balanced on these factors. For the three-year-old age groups, children of parents who had not completed high school were underrepresented by about five percent, with a corresponding overrepresentation by children with parents who had four or more years of college. The northeastern and north central regions were underrepresented by about three and four percent, respectively, with a corresponding overrepresentation in the southern (six percent) and western (two percent) regions.

The percentage of children passing each concept (sum of two items) is presented by age interval in the manual. The age intervals begin with three months (3-0 to 3-3), proceed by six-month intervals (3-3 to 3-9, 3-9 to 4-3, 4-3 to 4-9) and end with 4-9 to 5-0. Raw scores can be converted into percentiles for each age interval.

Examination of the percentage of children passing each concept, as well as the means and standard deviations of scores for each age group, indicates that the **Boehm-PV** measures well in the 3 to 4 age range. In the 4-3 to 4-9 and 4-9 to 5-0 age intervals, only the most difficult concepts (after, shortest, together, before and farthest) appear to significantly differentiate scores above the 25th percentile (i.e., the upper 75% of the scores).

Despite the care taken with the representativeness of a national sample, there are important factors beyond the limited numbers of children within each age interval that limit the interpretation of age-related norms. All the information used for test development and norming relates to children who have had some pre-school experience. No information is provided on the extent or the academic nature of that experience. No information is provided contrasting children of the same age who have not had pre-school experience. While the sample was balanced on sex within each age group, no mention is made of sex differences in performance.

Reliability: Overall, the evidence of reliability is judged to be good.

Internal consistency was measured separately for each age interval by two different methods, resulting in reliablity coefficients ranging from .91 to .80, with averages of .88 and .85 across all age intervals for the different methods.

Test-retest reliability was examined in a subsample of 78 children, ages 3 1/2 to 4 1/2, from the standardization sample. These children were administered the **Boehm-PV** twice, approximately one week apart. Reliability coefficients were .94 and .87 for ages 3 1/2 and 4 1/2, respectively, with a total of .94, strong evidence for the stability of test scores *in these age groups*.

The test-retest reliabilities were high but limited to only two of the age intervals.

Validity: The validity of the **Boehm-PV** is rated good in terms of its use as a measure of the mastery of concepts related to the language of instruction. This rating is based on the excellent and extensive focus on content validity (including pilot testing), and the limited but strong evidence of a concurrent relationship with another measure of language ability.

Content validity. In terms of use as a mastery test, a pre- post-test of instructional effectiveness, or as a measure of readiness for standard classroom instruction, the match between the set of concepts addressed in the **Boehm-PV** and those taught or required in the specific program using the test is the ultimate measure of content validity. The specific content of the **Boehm-PV** was selected on the basis of the following criteria:

- a review of research literature regarding the order and age of acquisition of basic language concepts
- analysis of tape recordings of classroom "teacher talk" used at the pre-school and primary grade levels when instructing or conversing with children
- extensive item tryouts with more than 300 children, from a variety of backgrounds and enrolled in a variety of pre-school programs
- review by a panel of educational specialists for appropriateness of content, including artwork, and potential bias toward particular subgroups of the population.

Criterion-related validity. Two studies are presented as evidence of concurrent validity; i.e., comparing performance on the **Boehm-PV** with performance on tests measuring similar or comparable abilities. Twenty-nine children, mean age 3-10, were given the **Boehm-PV** and the **Peabody Picture Vocabulary Test (PPVT-R)**, with a resultant correlation of .63. Nineteen language-delayed children, mean age 4-4, were given the **Boehm-PV** and the **PPVT-R**, with a resultant correlation of .57.

Evidence of validity for the **Boehm-PV** is limited at this time; however, the research base for the original **Boehm Test of Basic Concepts (BTBC)** suggests that further evidence will be provided as the instrument becomes more widely used.

Utility:

The **Boehm-PV** is a relatively quick, easily administered test of the mastery of basic concepts. The format of the picture book and individual record form makes the the test particularly easy to administer and score. The concepts are presented clearly with simple and appealing illustrations.

The **Boehm-PV** is particularly appropriate for use with children ages 3-3 to 4-3. The overall level of difficulty for three-year-olds indicates that the test measures well across a range of individual differences. However, the necessity of testing all 52 items means that most children are going to fail a substantial number of items. It would be much better to arrange the items in smaller difficulty groupings so that an individual ceiling level could be established for each child. The **Boehm-PV** would not yield much information about the level of concept mastery among higher performing children above age 4-3. It would be better to use the **Boehm-R** with children of this age range, unless some level of deficit in concept acquisition is expected.

While the **Boehm-PV** has substantial validity for use as a mastery test as a basis for instructional planning, the user should bear in mind the cautions about interpreting developmental age norms. The research base and the **Boehm-PV** item statistics do indicate that there is a definite developmental component to the acquisition of basic concepts. However, there is also evidence of a strong component of individual differences in rate of development and experiential factors. The significance of these individual differences is reflected in the large standard deviations for scores, particularly in the younger age groups. While performance on the **Boehm-PV** may very well be an indication of a deficit in language acquisition, the norm-referenced scores **should not be used** as the sole or primary indicator of children's underlying ability to acquire basic concepts.

The **Boehm-PV** would benefit from a broader standardization study, taking into account differences in parental and children's educational experience. Examination of item performance by SES or educational level, as was done for the **Boehm-R**, would be helpful in interpreting results for particular local populations.

Availability: The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204

Instrument: Bracken Basic Concept Scale - Diagnostic (BBCS, 1984)

Author: Bruce A. Bracken, Ph.D.

Purpose: The author's purpose for the full scale diagnostic instrument is to provide an in-depth assessment of a child's mastery of basic concepts to be used for individual and group instructional planning. The broad range of concepts addressed in the **BBCS** includes relational concepts (e.g., position, size), as well as labels for colors, shapes, textures, letters, and emotional states. [The **BBCS** screening forms are reviewed separately in the section on Screening Instruments.]

Description: The **BBCS** consists of 258 items, individually administered in a multiple choice format. [A given child takes only a portion of the 258 items, determined by individualized starting points, basals and ceilings.] The **BBCS** is intended for pre-school and primary children ages 2-6 through 7-11, and requires approximately 20-30 minutes when administered by a trained examiner. The eleven subtests and the items within each of subtests are arranged in increasing order of difficulty in the spiral bound stimulus manual and on the individual record form. The back of the stimulus manual folds out to become an easel which is placed facing the child and examiner for test administration.

The examiner begins each item saying "*Show me ...*" followed by the item stem as printed in the record form. Once it is clear that the child understands the task, just the item stem can be read (e.g., "*Which animal is big?*"). The child responds by pointing to (or saying the number of) one response choice. For the majority of items, there are four choices, arranged 2 x 2 on a page in the stimulus manual. The examiner must monitor the child's eyes to see that the child is looking at all the choices before responding. Standard instructions for administration should be read exactly as printed in the manual (initial instructions) and on the record form (item instructions).

The basic concepts addressed by the **BBCS** have been grouped into eleven subtests based on distinct conceptual categories. These categories and examples of item contents are listed below.

I.	*Color*	10 items -	color names
II.	*Letter Identification*	10 items -	5 upper, 5 lowercase letter names
III.	*Numbers/Counting*	14 items -	identify zero to nine objects, identify numerals 6 - 9
IV.	*Comparisons*	7 items -	"which fruits are **different**," "which boxes are **not the same**"
V.	*Shapes*	20 items -	basic one-, two- and three-dimensional shapes, concepts of "**in a line**," **in a row**"
VI.	*Direction/Position*	55 items -	behind, between, toward, right
VII.	*Social/Emotional*	29 items -	identify emotions, male/female
VIII.	*Size*	16 items -	big, tall, shallow, light
IX.	*Texture/Material*	24 items -	hard, shiny, cold
X.	*Quantity*	38 items -	full, many, whole, none, coin values
XI.	*Time/Sequence*	35 items -	finished, last, starting, over, seasons

The manual presents administration and scoring procedures, and suggestions for interpretation of results and instructional planning, as well as tables for scoring and detailed technical information about the test development.

In order to limit testing time and to test children at an appropriate difficulty level, starting levels, basals and ceilings are used. Every child begins with the first item in each of the first five subtests, and proceeds until three consecutive items are missed. The score on the first five subtests (called the "School Readiness Composite") is then used to determine the starting point (levels A through K) for the other six subtests. The items are administered in reverse order from this starting point until three consecutive items are passed (basal), and then forward until three consecutive items are missed (ceiling).

Scoring: All items are scored on a pass/fail (1/0) basis. Raw scores are converted into standard scores based on age in four-month intervals (2 years, 4 months through 7 years, 11 months). In addition to the total score, the **BBCS** provides subtest scores for the School Readiness Composite (the first five subtests combined) and each of the other six subtests.

The standard scores can then be converted into percentile ranks, stanines, or normal curve equivalents (NCEs) by reference to a second table. Raw scores can also be converted in "concept ages" by one month age intervals (total score) or two month age intervals (subtest scores).

Norms: The norms are rated as fair, in part due to the limited number of children per age interval. There are 17 four-month age intervals used for translating raw scores to standard scores. This would translate into approximate 65 children per interval *if* the ages were evenly distributed. There are no tables in the manual that indicate the actual age distribution of the normative sample, even by year of age.

The standardization sample consisted of 1,109 children. While the manual states that the sample was selected to represent the 1980 U.S. Census distributions of age, sex, ethnic group, geographic region, community size and socioeconomic status (SES), specific information is provided only for sex, ethnic group and geographic region. Some of the demographic information is clearly presented for the full scale and diagnostic scale standardization samples combined. It is not clear if and how the screening and the diagnostic scale samples overlapped.

The sample was representative in terms of percentages by sex and ethnic group. For the entire standardization sample, the southern and north central regions were under- and overrepresented by roughly 10%, respectively. No information is presented regarding the representativeness of the sample by community size within region. No information is presented to assess the representativeness of the sample by age.

The socioeconomic (SES) level of the sample participants was estimated on a site-by-site basis. While the manual states that an effort was made to represent low-, middle- and high-SES groups proportionately, no specific information is provided on what the actual representation was. No information is provided as to the pre-school experience of the standardization sample, except that Head Start, day-care, and public and private preschools were included..

The percentage of children passing each individual item (item difficulties) are presented only for screening tests A and B. Examination of the raw score to standard score conversion tables strongly suggests ceiling effects for most subtests after age 5-8. Only a few very difficult items differentiate scores in the upper one-third of the sample.

Reliability: The reliability of the **BBCS** is rated fair because of limited evidence. Evidence of reliability is presented for the age intervals much larger than those used in the norms and should be interpreted cautiously.

Stability over time (test-retest reliability) was examined in a study of only 27 children. The authors state that the age range of the sample was restricted, but no details of age or other sample characteristics are given. These children were administered the same form of the **BBCS** twice, approximately two weeks apart. Reliability coefficients ranged from .67 (size) to .95 for the subtest scores. Reliability coefficients were .98 for the School Readiness Composite and .97 for the total score.

Internal consistency reliability coefficients were presented by one-year age intervals and ranged from .47 to .96 for the subtests, with a range of .94 to .96 for the total test score. It would appear that the relatively low correlations for some subtests (particularly Size) were probably due to the ceiling effects restricting the range of scores.

Validity: Evidence for the validity of the **BBCS** is rated good.

Content validity: The author reviewed the contents and test directions of 13 preschool and primary cognitive and achievement tests, as well as curriculum materials, to arrive at a comprehensive list of more than 330 basic concepts. This list was reviewed by teachers, counselors and school psychologists. Because children typically confuse opposite or related concepts when they are learning, distractors (response choices other than the correct one) which were opposite the correct response or closely related (e.g., top: bottom, side, front, back) were deliberately chosen to determine if the child was in an interim stage of concept development. Initial item analysis and ordering by difficulty was conducted on the basis of a pilot study with 50 children across the age range of 2-6 through 7-11.

After the standardization data was collected, some of the items were eliminated and the final order of items was established on the basis of an item analysis.

Criterion-related validity: No evidence is reported for predictive validity of the **BBCS**. A number of studies are cited showing moderate to high correlations (.68 to .88) between the **BBCS** and the **Peabody Picture Vocabulary Test-Revised**, **The Boehm Test of Basic Concepts**, the **Token Test for Children**, and the **Metropolitan Readiness Test**.

Construct validity: The **BBCS** differentiated scores, as expected, between a group of 17 deaf children and a matched hearing sample. The scores of the deaf children were approximately two standard deviations below the mean.

Utility: The **BBCS** is certainly the most comprehensive measure of basic concepts available. It has substantial content validity, a crucial factor for the results to be useful in guiding instructional planning.

Overall, the **BBCS** is easy to administer and score. It would be much easier to use if there were tabs clearly marking the subtests and the starting levels within each subtest. There are a few items which are unnecessarily busy, and it might take longer for the child to visually isolate the information needed to understand the concept.

The **BBCS** needs stronger evidence of reliability, particularly test retest reliability, as well as evidence of predictive validity. While the level of technical detail provided in the manual is commendable in many cases, there are important details missing, such as the size of groups within age intervals. The author also recommends some questionable practices in terms of over-interpreting the normative information.

The author presents a method for profiling performance on the separate subtests and determining relative strengths and weaknesses that is not supported by the normative information. The author describes at length the psychometric determination of whether subtests measure unique variance and ends up with a significant number of cautions about subtest interpretation. It is difficult for someone not trained in psychometrics to understand the information provided, and the data is problematic enough that it is questionable whether subtest interpretation should be promoted. There is no justification of what utility subtest profiles have.

The author recommends the use of "concept age" scores as more readily interpretable than percentile ranks or standard scores. However, because the test covers concepts that are common in kindergarten and first grade curricula, the use of such scores is extremely questionable. It is conceivable that some children in the normative sample may have been in school one year longer than other children of exactly the same age. Therefore, the concept age averages the performance of these children.

Availability: The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204

Instrument: Cognitive Skills Assessment Battery (CSAB, 1981)

Author: Anne Boehm and Barbara Slater

Purpose: The authors' purpose is to provide an instrument to assess the mastery of cognitive and motor skills as an aid for instructional planning in prekindergarten and kindergarten programs.

Description: The **CSAB** includes 64 items in a format which combines oral, written and performance tasks. It is designed for children ages 3 to 6. The **CSAB** is individually administered and requires approximately 20-25 minutes to administer and score.

The items are administered orally by the examiner with the use of a card easel. The two-sided easel presents the information needed for each item on the side facing the child, while item instructions and scoring procedures face the examiner. The drawings are clear and represent a racial/ethnic mix of children. Some materials for administration need to be provided by the examiner, such as the eight blocks for the number knowledge task and a watch with a second hand.

Responses are recorded on the pupil response sheet which is also used by the child for printing his/her name and the visual-motor coordination tasks. A class record sheet, summarizing results from individual pupil response forms, can be used to record the profile of responses for each child and the class as a whole.

The manual contains instructions for administration and scoring, as well as field test comparative data. Cautions about interpretation of results are also presented. The manual also provides a discussion of test development and technical quality as well as suggestions for instructional planning.

The **CSAB** covers performance in five broad goal areas. The following are examples of specific item content for each area.

Orientation toward one's environment:	basic information, identification of body parts
Discrimination of similarities and differences:	color, shape, symbol, auditory and visual
Comprehension and concept formation:	number knowledge, information from pictures, story comprehension, multiple directions, vocabulary, letter naming
Coordination:	large muscle, visual-motor
Memory:	immediate, delayed recall, picture recall

Scoring: Depending on the specific task, items are scored on a pass/fail basis or by level of competence according to scoring criteria provided in the manual. Level 2 is full competence, Level 1, partial and Level N is complete lack of competence. After testing is completed, the examiner rates the child with a four-point scale on task persistence, attention span, body movement and attention to directions. Rather than a total score, a profile is obtained of performance on each task within each area.

Norms: Overall the normative information is judged to be fair.

The **CSAB** is not standardized but information about performance is presented for a field-test sample. The sample included 860 children tested in the fall and 558 children tested in the spring of 1980. The sample was judged to be broadly representative of the characteristics of the U.S. population with regard to geographic distribution, urban/suburban/rural community type and lower and middle socioeconomic level (SES).

Two kinds of comparative data are provided for fall and spring testing times. The percentage of children responding correctly is presented for each item by grade and separately by lower and middle SES levels. A ceiling effect for many items is indicated by more than 90% of the children answering correctly.

Reliability: The reliability of the **CSAB** is rated fair due to lack of evidence.

The author presents evidence of *stability* of test scores over a two to three week interval. There was 80% agreement overall for 16 prekindergarten children, and 85% agreement overall for 32 kindergarten children. Interrater reliability was 40% at the prekindergarten level and 79% for kindergarten children. These results should be interpreted with caution because of the small sample size.

Validity: Evidence for the validity of the **CSAB** is rated fair.

Content validity: The content was determined by a review of curricular materials and existing tests, the research literature, teacher interviews, classroom observations and field testing.

Utility: For the purposes of informal assessment to guide instructional planning, the user is the best judge of content validity. The ultimate value of this test is based on the curriculum match and the utility individual teachers see in the information provided by the **CSAB**.

The **CSAB** should not be treated as a norm-referenced test and the results should not be used to identify children or as a primary determinant in an important decision-making process.

Availability: Teachers College Press, Teachers College, Columbia University, New York, New York, 10027.

Instrument: The Lollipop Test: A Diagnostic Screening Test for School Readiness - Revised (1989)

Author: Alex L. Chew, Ed.D.

Purpose: The author's purpose is to provide a brief criterion-referenced "screening" instrument to identify children who will need additional instruction in readiness activities to obtain maximum benefit from their kindergarten and or first grade experience, for instructional planning and evaluation. The test was designed to be interesting to children of varying socio-economic backgrounds, inexpensive and amenable to local norming.

Description: **The Lollipop Test** consists of 49 items which are grouped into subtests as follows:

Test 1: Identification of colors, shapes and copying shapes
Test 2: Picture description, position, and spatial recognition
Test 3: Identification of numbers and counting
Test 4: Identification of letters and writing

The Lollipop Test is individually administered, requiring 15 to 20 minutes for administation and scoring. It requires only a brief orientation period for the novice examiner. The kit includes a combination Administration and Scoring Booklet and a set of seven spiral-bound stimulus cards. The subject matter is familiar and appealing to children; the illustrations are bright and clear. The eight-page Administration and Scoring Booklet has clear, standardized instructions, although the red print on bright yellow can be somewhat hard to read. This manual also includes general instructions and general interpretation guidelines.

The Developmental and Interpretive Manual for **The Lollipop Test** provides information about test development, descriptions of a number of validity studies, further guidelines for interpretation of results, and a discussion of developing local norms.

Scoring: Each item is scored on a pass/fail basis with one point assigned to all but copying shapes, which are scored with two points for each of three shapes copied correctly. The Administration and Scoring Booklet presents scoring criteria for each item type. **The Lollipop Test** is scored as a criterion-referenced test with total raw scores for each subtest and a total score.

Norms: National norms have not been established. However, because means and other descriptive information has been provided for a number of study samples, the normative information is rated fair.

It is suggested that school systems develop local norms if they wish to utilize norms instead of the suggested criterion-referenced approach. Instructions for developing local norms are included in the manual.

Descriptive statistics including means, medians, standard errors of the mean, and standard deviations are presented for the validation study sample of 69 Head Start and kindergarten children tested at a mean age of 70 months. Five suggested score ranges for readiness (below average, low average, average, high average and above average readiness) are presented. The 1989 revision of the manual contains descriptive statistics and interpretive score ranged for two additional study samples. 293 children (30% were black), tested at a mean age of 74 months were followed through grade 4. One hundred twenty-nine children (24% were black), tested at a mean age of 62 months, were followed through grade 1.

There is no rationale presented for the suggested score ranges, and the data do not appear to support such fine distinctions (i.e., barely more than one standard deviation difference between the highest and lowest score ranges). The descriptive statistics suggest that there may be a ceiling effect for **The Lollipop Test**, particularly when children are tested at the end of kindergarten. For groups of children tested in the spring before grade 1, the mode (most frequently occurring score) was the total possible for all subtests. The range of scores was quite high, however, suggesting that the test measures better at the lower end of the skill range.

Reliability: Reliability of **The Lollipop Test** is rated fair, because the evidence is limited.

The internal consistency (KR-20) reliablity ceofficient was .93. Test-retest reliability was not reported.

Validity: Evidence for the validity of **The Lollipop Test** is rated good.

Content validity: **The Lollipop Test** is one of the most technically rigorous tests in terms of the establishment of content validity. Individual test items were chosen on the basis of established predictive relationships with achievement and their presence on most readiness tests. Further, factor analytic studies were used to reduce the number of items to those that measure unique aspects of readiness (i.e., to reduce redundancy).

The content of **The Lollipop Test** is based on specifiable and teachable units of behavior that teachers consider important for children entering school. The author spent many years assessing young children as a school psychologist.

Construct validity: The manual begins with an interesting discussion of the "concept and theory of readiness." A complete review of the literature on which test development was based can be found in the author's dissertation. The idea of a short but valid readiness test was based on factor analytic studies of readiness. Correlations between subtests and the total **Lollipop Test** were high (.75-.89), showing that the test was measuring a consistent construct. The subtests were based on factor analysis.

Criterion-related validity: The sample for the initial validation study consisted of 69 kindergarten and Head Start students, with a mean age of approximately 5-10. The sample was nearly equally divided with regard to sex and race. The children had been enrolled in kindergarten for an average of seven months. The children were administered the **Metropolitan Readiness Test** (**MRT**) concurrently with **The Lollipop Test**, as well as having readiness skills rated by teachers on a scale developed for this study. The correlation between **The Lollipop Test** and the **MRT** was .86; between **The Lollipop Test** and teacher ratings, .58.

The 1989 revision of the manual describes two additional, longitudinal validiation studies. A sample of 293 children (30% were black), tested at a mean age of 74 months with both **The Lollipop Test** and the **MRT**, were followed through grade 4. Scores on **The Lollipop Test** predicted teacher assigned grades and performance on the **Stanford Achievement Test (SAT)** in grades 1, 3 and 4 as well as the much longer **MRT**. Correlations with the **SAT** ranged from .75 for reading and .72 for math in grade 1 to .40 for both reading and math in grade 4. Correlations with teacher assigned grades ranged from .54 for reading and .49 for math in grade 1 to .43 for reading and .30 for math in grade 4.

A sample of 129 children (24% were black), tested four months prior to kindergarten entry (mean age of 62 months) with both **The Lollipop Test** and the **DIAL**, a screening test, were followed through grade 1. Again the predictions from the shorter **Lollipop Test** were almost identical to the longer test. Classification analysis was not reported, however, to support the use of **The Lollipop Test** as a screening device.

Utility: **The Lollipop Test** is a brief, appealing, easily administered assessment of readiness skills. A great deal of effort went into chosing valid items for this test. The validity has been supported by longitudinal studies. Although this test has the current key word "diagnosis" in its title, the manual makes it clear that it does have the excess meaning applied to Meisel's (1985) definitions of diagnostic tests.

The author does suggest that **The Lollipop Test** can be used to identify children in need of further evaluation. The only evidence of its validity for screening is correlational evidence that it predicts later achievement as well or better than the **DIAL**. This is not sufficient evidence to support use of **The Lollipop Test** as a screening device, particularly since there are no national norms.

The 1989 revision of **The Lollipop Test encompasses expansions of the Developmental and Interpretive Manual to include further validity studies, and greater flexibility in the Administration and Scoring Booklet to allow pre- and post-testing on the same booklet. The test items themselves were not changed so the revision does not affect the significance of the validity studies.**

Availability: Humanics Limited, P.O. Box 7447, Atlanta, Georgia 30309

Instrument: School Readiness Survey, Second Edition (SRS, 1975)

Authors: F.L. Jordan and James Massey

Purpose: The author's purpose is to provide an instrument to help school personnel involve parents of preschool children in evaluating their child's developmental level (specifically in terms of skills needed in kindergarten, i.e., readiness skills) and in preparing the child for kindergarten. The **SRS** was designed to be administered and scored by the parent with school supervision. There are suggestions for fostering development of specific skills with simple tasks that can be accomplished in the home.

Description: The **SRS** includes 69 items which require primarily verbal and pointing responses from the child. The **SRS** is individually administered by the parent. Clear instructions, including exact wording of questions, are presented in the test booklet. Instructions are printed on alternate pages facing the opposite direction as the test items, so the parent can read the instructions with the item stimuli or response form facing the child. The manual suggests using a strip of paper as a marker to help the child keep his/her place. The **SRS** consists of seven sections as follows:

Number Concepts:	7 items	counting objects and by rote
Discrimination of form:	11 items	visual discrimination of simple forms, letters or objects
Color naming:	7 items	primary colors, plus green, orange, purple, and pink
Symbol matching:	4 items	visual discrimination of matching objects, letters or words
Speaking vocabulary:	20 items	naming familiar objects
Listening vocabulary:	4 items	identifying by gesture 12 familiar objects and categories
General information:	16 items	knowledge of name, age, address, and other aspects of the child's environment or common events; memory for a number series and sentence; and analogies

The additional General Readiness Checklist is a series of questions regarding the child's maturation and experiences that could not be directly tested.

Scoring: The parent is instructed to score as correct only those items the child actually performs correctly at the time of testing, even if he/she feels the child knows the correct answer. Scoring criteria are noted after the directions for each item. A brief section for parents on intepretation of scores is printed in the test booklet, along with score ranges for each section and for the total survey that indicate "ready for school," "borderline readiness," and "needs to develop." Parents are urged to contact school personnel if they have any questions about the results. The importance of each skill area for specific learning tasks and child-friendly suggestions for building skills in each area are also presented in the test booklet.

Norms: The norms are outdated and rated fair.

The original standardization sample included 842 children (and parents) from 18 schools representing a wide range of SES levels in one school district. A restandardization was completed in May of 1975, involving 383 preschool children from 20 elementary schools.

Reliability: Reliability of the **SRS** is rated fair.

Test-retest reliability was examined with an administration in June and a retest in October. The average gain in score over the summer was five points. In one study with a sample of 32 children teachers administered the **SRS** both times. The reliability coefficient was .79. For a second group of 20 children, the parents administered the **SRS** in June and teachers administered it in October. The reliability coefficient was .64. On the average parents tended to rate their children from two to five points higher than did trained administrators.

Validity: Evidence for the validity of the **SRS** is rated fair.

Content validity: Individual test items were chosen on the basis of interviews with kindergarten teachers and analysis of evaluations used for kindergarten children (i.e., grading criteria). Items which could not be tested directly (e.g., responsibility, alertness to environment) were included on a checklist for the parents. Items which required professional training to administer or score were discarded. The trial edition was piloted with 100 parents.

Criterion-related validity: The 383 children tested for the restandardization of the **SRS** in May of 1975 before kindergarten entry were followed up one year later with teacher ratings of school progress. The correlation between SRS scores and teacher ratings was .62. Correlations with SRS subsections ranged from .52 for number concepts and .50 for general information to .36 and .37 on the listening vocabulary and color naming, respectively.

Utility: Although the manner in which the manual discusses the use of the **SRS** in the context of school entrance decisions is dated, it should not be allowed to detract from the value of the **SRS** for the more "legitimate" uses as a guide for individualized program planning and, more important, as an effective communication device with parents. The **SRS** should not be used for screening.

Availability: Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, CA 94306

Instrument: Test of Early Language Development (TELD, 1981)

Authors: Wayne P. Hresko, D. Kim Reid, and Donald D. Hammill

Purpose: The authors' purpose is to provide a well-constructed, standardized measure of early spoken language based on current theoretical perspectives. It is intended for use to identify children who are in need of more extensive, clinical evaluation, to document children's progress in language and to suggest instructional practice.

Description: The **TELD** is an individually administered test consisting of 38 items requiring oral and pointing responses. The kit includes picture cards which are presented to the child for some items. It is untimed but the authors report it usually can be administered in 15-20 minutes and scored in an additional 10 minutes. Because of the range of ages tested, separate starting points have been established for each year of age, as well as basal and ceiling points of 5 items. There are standard instructions for each item.

The **TELD** addresses two dimensions of Language - content (encode, decode meaning) and form (syntax, morphology, phonology) - in receptive and expressive modes. Examples of the content for each of these categories are:

Content, receptive:	Show me the ball (blanket, cup).
Content, expressive:	What is your favorite TV show? Billy was tired. He hadn't taken a nap. What do you think he would say to his mother? (What would you say?)
Form, receptive:	Show me "The car hit the truck."
Form, expressive:	Say each word after me. Say "fine" ("blue", "seven"). I'm going to say some sentences. Say them exactly as I say them. "The girl likes walking by herself."

Scoring: Individual items are scored on a pass/fail basis (1/0) according to scoring criteria explicitly stated for each item. The total test score and the analysis of the type of items passed and failed are the primary guides in inferences about the child's language abilities. Test performance is reported in terms of three kinds of normative scores including the Language Quotient (deviation standard scores), Percentiles and Language Ages (based on the average score of children within each six month age interval). Instructions and cautions about the interpretation of scores are given in the manual. The suggested interpretation of **TELD** scores is primarily based on the Language Quotient which is interpreted in a similar fashion as an Intelligence Quotient.

Norms: The normative information for the **TELD** is rated good.

The standardization sample included 1184 children from eleven states and one Canadian province. Except for slight overrepresentations of the Southern geographical region and of "White-Collar" parents, the sample was representative of the United States population as reported in 1979.

Reliability: The reliability of the **TELD** is rated excellent.

Internal consistency of test items was examined during test construction. The coefficients ranged from .87 (age six) to .92 (age three), with a mean of .90. Reliablity coefficients for the *test-retest* performance of 177 children, ages three to seven ranged from .72 to .87; the coefficient for the total group was .90. The Standard Error of Measurements for each age group rounded to two raw score points, indicating a high level of confidence in raw scores.

Validity: Evidence for the validity of the **TELD** is rated fair.

Content validity. Item content selection was theoretically based and is described in the manual. The final 38 items were selected from an initial 370 on the basis of two pilot studies (200 and 100 children from two separate geographic regions) and detailed item analysis.

Construct validity. The internal consistency reliabilities ranged from .87 to .92 (for three year olds), indicating that the items address the same trait. The **TELD** shows a clear differentiation of scores by age supporting the authors' contention that the **TELD** is measuring a developmental trait. It is related to measures of intelligence, reading and school readiness supporting the authors contention that it taps abilities influenced by the cognitive/thinking process. A small sample of children already diagnosed as "communication disordered" (ages 3-0 to 6-6) scored nearly two standard deviations below the mean on the **TELD**, lending some support to its ability to identify children with independently confirmed communication problems.

Criterion-related validity. Scores from the **TELD** were modestly correlated with those of the Reading Subtest of the **Metropolitan Achievement Test** (6-year olds, .34), the Composite Score from the **Test of Reading Comprehension** (7-year-olds, .55), the **TELD** (3- to 6-year olds, .82), the Matching and Alphabet subtests of the **Metropolitan Readiness Test** (6-year-olds, .42, .54) and the **Slosson Intelligence Test** (5-year-olds, .78). No evidence of validity is presented for four-year-old populations.

Utility: The **TELD** is a brief, easily administered test of early language skills, both expressive and receptive. The manual is clear and explicit and the normative information is of good quality and appears appropriate across the entire age range. Three-year-old children may not get very far into the test (mean score = 9, sd 6), but the items are not intimidating and the test shows good reliability at that age. The authors indicate that the test is most effective with four, five and six-year-old children.

There is not sufficient evidence to support the validity of the **TELD** as a screening measure. It would greatly support the validity to have more research on the **TELD**'s ability to identify at-risk children.

There is a Spanish version of the **TELD**, the Prueba del Desarrolo Inicial del Lenguaje (**PDIL**), standardized on a sample of Spanish speaking children living in Mexico, Puerto Rico and the United States.

Availability: Pro-Ed, 5341 Industrial Oaks Boulevard, Austin, TX 78735.

Instrument: Test of Early Mathematics Ability (TEMA, 1983)

Authors: Herbert P. Ginsburg and Arthur J. Baroody

Purpose: The authors' purpose is to provide a test of early "informal" mathematical thinking which serves as a foundation for the "formal" mathematical skills taught in school. Knowledge of strengths and weaknesses in this foundation can be used in planning effective educational strategies. The scope of the **TEMA** goes beyond informal mathematics to include knowledge of formal rules, principles and procedures. This test can be used to document progress in learning arithmetic and to identify children who are significantly behind or ahead of their peers in the development of mathematical thinking.

Description: The **TEMA** is an individually administered test designed for children ages 5 to 8. The 50 items require a range of oral and performance responses from the child, including calculating by manipulating small objects and by writing on paper (older children). The kit includes response cards which are presented to the child for some items. The examiner needs a supply of small, countable objects.

The **TEMA** is untimed but the authors report it usually can be administered in about 20 minutes. Because of the range of ages tested, separate starting points have been established for each year of age, and basals and ceiling points of five items are used. There are standard instructions for each item. Test content covers the following areas:

Informal Mathematics

- *Concepts of Relative Magnitude*. Items tapping this concept begin with the concept of "more", then focus on the ability to judge relative distances between numbers on a mental numberline.
- *Counting*. This skill is the most heavily represented in the 23 items measuring informal mathematics. Item content includes rote counting, counting backwards and counting objects (enumeration).
- *Calculation*. Items tapping this skill range in difficulty from adding concrete objects to mental addition and subtraction.

Formal Mathematics

- *Knowledge of convention*. Items measure the fundamental skills of reading and writing numbers.
- *Number facts*. Items include simple addition and subtraction problems that must be answered quickly to indicate knowledge rather than on-the-spot solution.
- *Calculation*. Items measure the accuracy and process of addition and subtraction. The child is asked to talk aloud as the problem is being solved.
- *Base-Ten concepts*. Money problems are used to test base-ten concepts as well as place-value items.

Scoring: Individual items are scores as pass/fail, even when multiple steps are required to arrive at the answer. Test performance is reported in terms of three kinds of normative scores including the Math Quotient (deviation standard scores), Percentiles and Math Ages (based on the average score of children within each six-month age interval).

Norms: Normative information is rated fair.

The standardization sample included 617 children from 12 states. In terms of the information provided, the sample is fairly representative of the the United States population as reported in 1981. There is an underrepresentation of the Northeastern geographical region and an overrepresentation of city residence with rural underrepresentation. The most questionable feature is that 50% of the parental occupation category is unknown. This should be an important factor in judging the appropriateness of the norms, particularly when one of the expressed purposes of the test is to rank performance relative to "peers."

Reliability: Reliability of the **TEMA** is rated fair.

The coefficients reported for *internal consistency* of test items range from .87 (age three) to .93 (age seven). The reliablity coefficient for the *test-retest* performance of 71 four- and five-year-old preschool children was .94. No test-retest reliability is reported for school-aged children. While correlations were high, the evidence for reliability is base on small samples that do not cover the age range of the test.

Validity: Evidence for the validity of the **TEMA** is rated fair.

Content validity. Item selection was literature based and the final 50 items were selected from an initial pool of 96 on the basis of two pilot studies and detailed item analysis.

Construct validity. The **TEMA** shows a clear differentiation of scores by age. Scores on the **TEMA** for 62 four- and five-year-old preschool children had correlations of .66 with the Slosson Intelligence Test and .39 with the **TELD**. A study which found significant differences in performance on the **TEMA** between "high risk" and "normal" children supports the construct validity.

Criterion-related validity. Scores from the **TEMA** were correlated with the Math Calculation subtest of the Diagnostic Achievement Battery. The coefficients were .40 for a sample of 23 six-year-olds and .59 for a sample of 17 eight-year-olds. No evidence of predictive validity is presented for children younger than six.

Utility: The **TEMA** is a brief, easily administered test of early mathematics skills. While the age range of the test is listed at 4-0 to 8-11, the manual states that the test is too difficult for most four-year-old children, unless they are unusually gifted in math. The number of items appropriate for preschool or beginning kindergarten is very limited. There will be a second edition of the **TEMA** available in September 1989 which will have items appropriate for children as young as three.

The **TEMA** manual begins with an interesting research-based discussion of the nature of early mathematical thinking. However, items tapping the "informal" aspects of children's knowledge rely heavily on complex counting tasks. The evidence of reliability and validity tends to be limited. There is not sufficient evidence to support use of the **TEMA** as a screening measure.

Availability: Pro-Ed, 5341 Industrial Oaks Boulevard, Austin, TX 78735.

Instrument: Test of Early Reading Ability (TERA, 1981)

Authors: D. Kim Reid, Wayne P. Hresko, and Donald D. Hammill

Purpose: The authors' purpose is to provide a test of early reading (rather than reading "readiness") which can be used to document children's progress in reading and identify those children who are significantly behind their peers in reading development.

Description: The **TERA** is individually administered. The kit includes response cards which are presented to the child for each item. The test is untimed, but the authors report it usually can be administered in 15-20 minutes. The **TERA** is designed for children ages 4 to 8. Because of the range of ages tested, separate starting points have been established for each year of age. There are clear, standard instructions for each item. Testing is ended when the child misses five consecutive items.

The specific content of the **TERA** is described as follows:

Finding meaning in print. The majority of items address this component of early reading.

Specific item types include the following:

- Reading signs, logos and words frequently encountered in figural/situational contexts
- Relational vocabulary
- Discourse, including retelling a story, anticipating written language (e.g., on a birthday card), and a cloze task of comprehension during silent reading

The alphabet and its functions. Specific items types include letter naming, oral reading and proofreading (finding errors).

Print conventions. Specific item types include book handling, punctuation, left-right orientation, and the spatial presentation of a story on the page.

Scoring: Items are scored on a pass/fail basis. Test performance is reported in terms of three kinds of normative scores including the Reading Quotient (deviation standard scores), Percentiles and Reading Ages (based on the average score of children within each six-month age interval).

Norms: Overall, the norms are rated good.

The standardization sample included 1184 children from 11 states and one Canadian province. Except for slight overrepresentations of the southern geographical region and of "white-collar" parents, the sample was representative of the United States population as reported in 1979.

Reliability: Reliability is rated excellent.

Internal consistency of test items was examined during test construction. The coefficients range from .87 (age three) to .93 (age seven). Reliablity coefficients for the *test-retest* performance of 177 children, ages three to seven ranged from .85 to .94; the coefficient for the total group was .97.

Validity: Evidence for validity is rated fair because it is limited.

Content validity: Item selection was literature based (described in the manual) and the final 50 items were selected from an initial 270 on the basis of two pilot studies and detailed item analysis.

Construct validity: The **TERA** shows a clear differentiation of scores by age. It is related to measures of intelligence, language and school readiness supporting the authors contention that it taps abilities influenced by the cognitive/thinking process. A small sample of children already diagnosed as "learning disability/reading disordered" scored more than one standard deviation below the mean on the **TERA**, lending some support to its ability to identify children with reading problems.

Criterion-related validity: Scores from the **TERA** were correlated with those of the Reading Subtest of the MAT (6-year olds, .66) and the Composite Score from the Test of Reading Comprehension (7-year-olds, .52).

Utility: The **TERA** is a brief, easily administered and psychometrically sound test of early reading skills. The authors found that the test is much too difficult for children under four and recommend that it not be used below that age although test development and normative information is presented from age three. Even at age five the median item difficulty (percent of children passing) is only 34. The authors do not present any evidence of predictive validity between the ages of four and six. A median item difficulty of 91 suggests a possible ceiling effect at age seven.

The **TERA** manual begins with a research-based, theoretical discussion of the nature of early reading. It describes the characteristics of the test (see Description) in terms of many interesting pre-reading skills. However, very few of these pre-reading skills are actually in the test. Six of the first 15 items deal with letter knowledge and after item 15 the majority of the items require reading. The authors do state that their purpose is to focus on reading rather than reading readiness; however, this limits the usefulness of the **TERA** in an ECE context.

There is no evidence that the **TERA** is valid as a screening measure

Availability: Pro-Ed, 5341 Industrial Oaks Boulevard, Austin, TX 78735.

Instrument: Test of Language Development-2, Primary (TOLD-2 Primary, 1988)

Authors: Phyllis L. Newcomer and Donald D. Hammill

Purpose: The authors' purpose is to provide a norm-referenced instrument to assess the mastery of expressive and receptive language skills. It is intended to identify children who are significantly behind their peers in language proficiency, to determine children's specific strengths and weaknesses in language skills, to determine progress in language as a consequence of intervention programs, and as a measure in research studies.

Description: The **TOLD-2 Primary** includes 175 items requiring a variety of verbal responses from children as well as pointing to correct pictures in a multiple choice format. It is designed for children ages 4-0 to 8-11. The **TOLD-2 Primary** is individually administered and requires 30-60 minutes to administer and score, depending on the age and ability level of the child. An abbreviated 55-item version for large-scale screening or research purposes consists of the two subtests most strongly correlated with the total score.

Many items are administered orally. The kit includes picture cards which are presented to the child for some items. The subtests should be administered in the same order that was used when the test was standardized. All subtests begin with the first item and testing is stopped at a ceiling of five items missed in succession. There are standard instructions for each item type in the manual which should be read exactly as printed. For convenience the directions are also printed on the record sheet.

The **TOLD-2 Primary** is based on a two-dimensional model of language made up of linguistic features (phonology, syntax, morphology, semantics) and linguistic systems (listening/receptive, and speaking/expressive). The chart below illustrates this model, showing the relationships among the primary concepts and the subtests.

	Linguistic Systems	
Linguistic Features	Listening (Receptive Skills)	Speaking (Expressive Skills)
Semantics	*Picture Vocabulary* *(PV)* 25 items	Oral Vocabulary (OV) 20 items
Syntax	Grammatic Understanding (GU) 25 items	Sentence Imitation (SC) 30 items
		Grammatic Completion *(GC)* 30 items
Phonology	Word Discrimination (WD) 25 items	Word Articulation (WA) 20 items

Scoring: Individual items are scores on a pass/fail basis according to detailed scoring criteria presented with examples in the manual. Raw scores are transformed into percentiles and standard scores via tables in the manual appendix. Percentiles and standard scores can be recorded for each separate subtest. Standard scores can then be summed across various groupings of subtests to arrive at "quotients" for each of the following composite constructs:

Spoken Language:	all subtests
Listening:	Picture Vocabulary, Grammatic Understanding, Word Discrimination
Speaking:	Oral Vocabulary, Sentence Imitation, Grammatic Completion, Word Articulation
Semantics:	Picture Vocabulary, Oral Vocabulary
Syntax:	Grammatic Understanding, Sentence Imitation, Grammatic Completion
Phonology:	Word Discrimination, Word Articulation

Because reliablity and validity have been established for each subtest, subtests may be used independently if a complete battery is not needed. For a quick screening of large numbers of children (to identify those who may have a language problem), standard scores from Picture Vocabulary and Grammatic Completion can be summed for a separate quotient which provides an estimate of the Spoken Language Quotient (SLQ). These two subtests were chosen because this combination yielded the highest correlation with the SLQ.

A table is given to convert the standard scores into NCE -, T-, z-scores, or stanines. Thorough descriptions of what the subtests measure, instructions and cautions about the interpretation and sharing of scores and quotients are given in the manual. The authors caution against the use of age equivalent scores but do provide a conversion formula for use when legislative or school policies require such scores. An optional software scoring system is available.

Norms: Overall the norms are rated excellent.

The standardization sample included 2,436 children from 29 states and one Canadian province. The sample was representative of the United States population (as reported in 1985) in terms of sex, race, place of residence (city/rural), geographical distribution and occupation of parents. In addition, means and standard deviations of samples from a variety of research studies are presented in the manual for comparison purposes.

The standard score norms were initally derived from the cumulative frequency distribution of raw scores for each six month age interval. Where differences between means were one raw score or less, the data for intervals were combined (Oral Vocabulary, Grammatic Understanding, Word Discrimination). In the case of Word Articulation, ages 5-6 to 6-11 were combined, possibly due to a ceiling effect.

Instructions for and cautions about constructing local norms are presented in the manual.

Reliability: Evidence for reliability is rated excellent.

Internal consistency was examined, separately by age, for each subtest and for composite scores. The coefficients for the subtests ranged from .81 to .95; coefficients for the composite scores ranged from .88 to .97. Coefficients for the overall Spoken Language Quotient ranged from .96 to .97, and coefficients for the short-form estimate of the SLQ .91 to .93 *Internal consistency* of test items was also examined for a sample of 37 children diagnosed as having disorders in oral communication. The coefficients for the subtests ranged from .80 to .89; the coefficient for the total scores was .95.

Reliablity coefficients for the *test-retest* performance of 21 children ranged from .74 to .95 for individual subtests; the coefficients for the composite scores from .80 to .93, with .94 for the SLQ. A separate sample of 59 children yielded coefficients that ranged from .86 to .98; the coefficients for the composite scores were all .98, except for a .99 for the SLQ. In both these studies the effects of the wide range of ages were statistically controlled.

Validity: Evidence for the validity of the **TOLD-2 Primary** is rated good.

Content validity. Item content selection was theoretically based (thoroughly described in the manual) and guided by well-known tests of the separate constructs addressed. Items were selected for each subtest on the basis of at least two pilot studies and detailed item analysis. The meaningfulness of the specific subtests chosen for the **TOLD-2 Primary** in measuring the features and systems of language was validated by a survey of 100 professionals including authors, reviewers for journals, college professors and school personnel involved in language assessment.

Construct validity. The **TOLD-2 Primary** shows a clear differentiation of scores by age in a number of separate studies, supporting the authors' contention that it is measuring developmental abilities. The subtests are moderately related to one another as would be predicted from tests measuring various aspects of language.

The **TOLD-2 Primary** is related to measures of intelligence, reading, writing, school readiness and general achievement, supporting the authors' contention that a test of spoken language should be related to tests of school achievement and readiness.

In a number of studies with children diagnosed as retarded, or as having speech/language or academic problems, the results for the **TOLD-2 Primary** were marked below that of the standardization population, supporting the claim that the **TOLD-2 Primary** can differentiate such groups from typical children. Eighteen studies are summarized in the manual.

Criterion-related validity. No evidence of predictive validity is available for the **TOLD-2 Primary** as yet. In terms of concurrent validity there is a wide base of research based on the earlier, but very similar version of the **TOLD-2 Primary**.

Scores from the separate subtests of the **TOLD-2 Primary** were correlated with scores from other widely accepted tests which addressed the same construct. These correlations were moderate to strong (.49 to .86) for children in three age groups (4, 6 and 8). A number of separate studies substantiate the validity of the **TOLD-2 Primary**.

The short form (Picture Vocabulary and Grammatic Completion subtests) has validity as a screening measure in terms of its strong relationship with the total **TOLD-2 Primary** score. However, no studies have examined the validity of referral classification (i.e., sensitivity, specificity) with either the **TOLD-2 Primary** or the short version.

Utility: The **TOLD-2 Primary** is an easily administered, thorough, well documented and psychometrically sound test of a broad range of language skills. The normative information is of good quality and appears appropriate across the entire age range. There is strong evidence of reliability and good evidence of validity as a measure of language skills. The **TOLD-2 Primary** is not designed to directly guide instructional planning; However the content is relevant to instruction and related to achievement measures. While the short form of the **TOLD-2 Primary** appears to be reliable, no evidence is available for the predictive validity of its use as a screening device.

Availability: Pro-Ed, 5341 Industrial Oaks Boulevard, Austin, TX 78735.

APPENDIX F

SUMMARY TABLE OF READINESS MASTERY INSTRUMENTS

Content and Key to Instrument Descriptors in Review Summary Tables

INSTRUMENT: *Instrument name, acronym, author(s), publication date and publisher. Indices of instruments by title and publishers' addresses are included after Appendix J.*

FOCUS: *Scope of content covered by the instrument.*

Broad: Includes three or more of the following categories of abilities:
Language, Speech, Cognition, Perception, Personal/Social, Perceptual-motor, Fine, Gross Motor Coordination
Academics: Includes many, but primarily academic skills
Specific Areas: Language, Literacy, Mathematics, Reading, Relational Concepts
(see "Content" for specific skills in each area)

AGE/GRADE: *Age or grade range covered by the instrument.*

ADM. TIME: *Time in minutes required for administration and initial scoring.*

FORMAT: *Description of test in terms of type of response required, format and materials, categories are not mutually exclusive*

Format: Group or Individual Administration
Multiple choice
Paper & Pencil (child marks or writes the answer)
Stimulus cards/easel
Manipulatives (e.g., blocks, sorting chips)
Response Mode : Teacher rating
Parent response
Observation of Child
Oral (verbal)
Pointing (implies multiple choice)
Performance (*fine/visual-motor*: copy, build, write, etc)
Motor (*gross motor*: hop, skip, jump, catch, etc.)

SCORES: *Types of scores available. No endorsement of the use of specific types of scores is implied* here.

Norm-referenced: Percentile, Percentile Rank
Age Equivalent / Grade Equivalent (Gr.Eq.)
Standard Score
Normal Curve Equivalent (NCE)
Developmental "Age", "Language Age", etc.
Quotient (Developmental, Language, etc.)
Criterion-referenced: Mastery levels
Raw score

CONTENT: *When the content covers a number of areas, the category name is used. When the content is more limited within a category, the specific areas are named.*

Basic facts: colors (primary), letters, numbers. shapes
Language: expressive, receptive vocabulary, fluency, syntax
Literacy: print functions & conventions, reading symbols
Relational Concepts: direction, position, size, quantity, order, time, categorization
Listening & Sequencing: follows directions, remembers story sequences, main ideas
Cognitive: problem solving, opposite analogies, memory, imitation
Perception: auditory, visual discrimination
Mathematics: count rote, with 1/1 correspondence, number skills
Motor: fine motor (holding a pencil correctly, buttoning, etc)
gross motor (hops, skips, throws)
visual-motor (copies shapes, builds blocks)
Self: knowledge of body parts (point or name)
social/emotional (peer & teacher interactions, attention span, etc.)
self help (buttoning, toilet, etc)
information (name, age, address, phone, birthdate)

NORMS: *Ratings on norming studies (value judgement implied)*

None: no normative information is given
Poor: some information but limited applicability
Fair: some standards of comparison (e.g., means of research sample)
Good: norms based on good sized, representative sample,
or lots of relevant information regarding appropriate populations for use
Excellent: norms based on a representative, national sample and relevant
information about applying norms or norm-referenced scores.

RELIABILITY: *Reliability ratings (value judgement implied)*

None: no reliability information is provided
Poor: all reliability coefficients (r) below .70
or an important type of reliability was not examined
Fair: at least one reported **r** is greater than .70; or **r** was
greater than .80 but evidence was limited in applicability
Good: total **r** is greater than .80; most subtests have **r** greater than .75
Excellent: several kinds of reliability reported; total **r** is greater
than .90; most subtest scores greater than .80

VALIDITY: *Validity ratings (value judgement implied)*

None: no validity information is provided
Poor: information is of very limited applicability
Fair: most important aspects of were addressed but evidence was
moderate or weak; or was strong but limited in applicability
Good: consistent evidenct of validity, or strong but limited evidence
of the type of validity most appropriate for the intended test use
Excellent: strong evidence and a base of research on the instrument

Summary Table of Instrument Characteristics: Mastery of Readiness Concepts

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Analysis of Readiness Skills Rodrigues, Vogler & Wilson, 1972 The Riverside Publishing Company	Academics (Limited)	Grade K	30 - 40	Individual or Group Adm. Paper & Pencil Multiple Choice	Letter Discrim & Naming Number names & Counting	Percentile	Poor Dated	Poor Limited	Poor Limited	Traditional concept of readiness skills
Basic School Skills Inventory-- Diagnostic (BSSI-D) Hammill & Leigh, 1983 PRO-ED	Broad	Ages 4 - 6	20 - 30	Individually Adm Teacher ratings Performance Oral	Language Literacy Mathematics Self/behavior	Percentile Standard	Fair	Fair	Poor	
Boehm Test of Basic Concepts - Revised (Boehm-R) Boehm, 1986 The Psychological Corporation	Relational Concepts	Grades K 1 - 2	30	Group Adm. Paper & Pencil	All areas of Relational Concepts	Total Raw Score Percentile	Excellent	Grade K Good Overall Fair	Grade K Excellent Overall Good	Class record form = Key Parent/teacher Conference Report form available
Boehm Test of Basic Concepts - Preschool Version (Boehm-PV) Boehm, 1986 The Psychological Corporation	Relational Concepts	Ages 3 - 5	10 - 15	Individually Adm Paper & Pencil	All areas of Relational Concepts	Total Raw Score Percentile	Fair	Good Limited	Good Limited	Class record form = Key Parent/teacher Conference Report form available
Bracken Basic Concept Scale - Diagnostic (BBCS-D) Bracken, 1984 The Psychological Corporation	Relational Concepts	Ages 2 1/2 to 8	20 - 30	Indiviudally Adm Multiple Choice Pointing or Oral	All areas of Relational Concepts	Standard Percentile Stanines NCE	Fair	Fair	Good	Exhaustive set of 258 concepts The use of "concept age" score is not recommended
CIRCUS ETS, 1972, 1979 CTB/McGraw-Hill	Academics	Grades Pre-K K & 1	30 per subtest	Group Adm Paper & Pencill Multiple choice	Perception Mathematics Language Cognition	Standard Percentile Stanine	Excellent	Good	Good Limited	Many subtests can be used spearately or in groups; Teacher Observation Instrumt avail
Cognitive Skills Assessment Battery (CSAB) Boehm & Slater, 1981 Teachers College Press	Academics	Grades Pre K & K	20 - 25	Individually Adm Stim. Card Easel Oral, Perform. Written	Concepts Perception Cognition Self	% Pass by item Means for area	Fair	Fair Limited	Fair	Fall & spring norms by SES level Behavior rating scale available

Summary Table of Instrument Characteristics: Mastery of Readiness Concepts cont.

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Gesell Preschool Test Haines, Ames & Gillespie, 1980 Programs for Education, Inc.	Broad	Ages 2 1/2 - 6	30 - 45	Individually Adm. Manipulatives Oral & Performance	Self Language Visual Motor	Age based success level by item	Poor Limited	None	Poor Limited	Reliability and validity have not been established
Gesell School Readiness Test aka **School Readiness Screening Test** (SRST), 1978 Programs for Education, Inc.	Broad	Ages 4 1/2 - 9 4 1/2 - 5	20 - 30	Individually Adm Manipulatives Performance Oral	Self Language Visual Motor	Age based success levels	Poor Limited Dated	None	Poor Limited	Clinical approach to scoring requires extensive training
The Lollipop Test Chew, 1981, 1989 Humanics LTD	Academics	Grades Pre-K & K	15 - 20	Individually Adm Pointing, Oral Copying	Basic Facts Relt.Concepts Copy shapes Math & Writing	Raw Scores Suggested Mastery Levels	Fair	Fair	Good	Attractively packaged Child & examiner friendly
Metropolitan Readiness Tests- Fifth Editon (MRT) Nurss & MacGauvan, 1986 The Psychological Corporation	Academics	Grades Pre-K K & 1	80 - 95	Group Adm. Paper & Pencil Multiple Choice Performance	Language Literacy Perception Mathematics	Raw Score Percentile Stanine Mast. levels	Excellent	Good	Good	Instructional Materials Parent/teacher Conference Report forms Behavior checklists
Preschool Inventory (PI) Caldwell, 1970 CTB/McGraw-Hill	Academics	Ages 3 - 6	15	Individually Adm Manipulatives Oral Motor Performance	Self Language Basic Facts Copy Forms	Percentile % Pass by item	Fair Dated Limited	Fair Limited	Fair	Clear SES differences Norm group all Head Start children available
School Readiness Survey. Jordan & Massey, 1976 (SRS) Consulting Psychologists Press	Academics	Grades Pre K	Untimed	Individually Adm by the Parent Multiple Choice Pointing, Oral	Basic Facts Perception Cognitive Vocab. & Self	Readiness Levels	Fair Dated	Fair	Fair	Effective communication device to discuss school readiness with parents
Tests of Basic Experiences Second Editon (TOBE 2) Moss 1979 CTB/McGraw-Hill	Academics	Grades Pre K K & 1	160 40 per subtest	Group Adm Paper & Pencil Multiple Choice	Language Mathematics Science Social Studies	Standard Percentile Stanines NCE	Excellent	Good Limited	Fair Limited	Optional 1 item/page books Fall, winter, spring norms Public & Catholic norms Practice Test

Summary Table of Instrument Characteristics: Mastery of Readiness Concepts cont.

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Test of Early Language Development (TELD) Hresko, Reid & Hammill 1981 PRO-ED	Language	Ages 3 - 7	15 - 20	Individually Adm Stimulus cards . Oral Pointing	Expressive Receptive Vocabulary Syntax	Percentile Lang Quot Lang Age.	Fair Limited	Excellent	Good	Well written, helpful manual
Test of Early Mathematics Ability (TEMA) Ginsburg & Baroody, 1983 PRO-ED	Mathematics	Ages 4 - 8+	20	Individually Adm Stimulus cards . Manipulatives Oral, Perform.	Quantitative Concepts Counting Calculation	Percentile Math Quot Math Age.	Fair Limited	Good Limited	Fair	New version coming in 1989 This version has limited utility for preK or beg. K
Test of Early Reading Ability (TERA) Reid, Hresko & Hammill, 1981 PRO-ED	Reading	Ages 4 - 8+	15 - 20	Individually Adm Stimulus cards . Oral Pointing	Wide range of Early Literacy Skills	Percentile Standard Lang Age.	Good	Excellent	Fair Limted	All new version for 1989 This version difficult below age 6
Test of Early Written Language (TEWL) Hresko, 1988 PRO-ED	Literacy	Ages 3 - 8	10 - 30	Individually Adm Stimulus cards . Writing, Oral Pointing	Range of Early Literacy Skills	Percentile Standard	Fair Limited Informtn	Good Limited	Poor Limited	Administration instructions tend to hurry child Norms do not acount for experiential differences
Test of Language Development - Primary (TOLD-2 Primary) Hresko, Reid & Hammill 1981 PRO-ED	Language	Ages 4 - 8+	30 - 60	Individually Adm Stimulus cards . Oral Pointing	Expressive Receptive Vocabulary Syntax	Percentile Standard Lang Quot. T- z- NCE	Excellent	Excellent	Good	Well written, helpful manual

APPENDIX G

REVIEWS OF OTHER EARLY CHILDHOOD INSTRUMENTS

Contents of Appendix G

Developmental Inventories

Page

1 **DP II** Developmental Profile II
9 **HNCAF** Humanics National Child Assessment Forms

Cognitive Maturity

5 **EOWPVT** Expressive One-Word Picture Vocabulary Test
7 **HFDT** Human Figures Drawing Test
11 **PPVT-R** Peabody Picture Vocabulary Test-Revised
14 **ROWPVT** Receptive One-Word Picture Vocabulary Test

Miscellaneous

13 Readiness for Kindergarten: A Coloring Book for Parents

Instrument: Developmental Profile II (DP II, 1980)

Authors: Gerald D. Alpern, Ph.D. (1972 original and 1980 revision), Thomas J. Boll, Ph.D. (1972 original), and Marsha S. Shearer, M.A. (1980 revision)

Purpose: The authors' purpose is to provide a comprehensive inventory of skills to assess development from birth to age nine. It was designed to cover motor, language, personal/self-help, social and intellectual development in a relatively short period of time by an evaluator who is not necessarily an expert. The authors suggest many "valid and appropriate" uses for the **DP II**: for determining eligibility for special education or related services; as a tool to develop individualized educational programs consistent with a child's strengths and weaknesses; as a measure of educational progress; and for pre- and post-testing in program evaluation.

Description: The **DP II** consists of 186 skills covering an age range of 0 to 3-1/2 years in six-month intervals and in yearly intervals thereafter to age 9. An individual profile of a child's "functional developmental-age level" is provided by classifying skills by age norms in five areas of development. Examples:

Physical Age: large and small muscle coordination, strength, stamina, flexibility and sequential motor skills
Self-Help Age: eating, dressing, household chores
Social Age: taking turns, playing with others, awareness of sexual identity
Academic Age: classification, knowledge of colors, counting, rhyming, drawing forms and persons
Communication Age: expressive and receptive communication skills, use and understanding of spoken, written and gestural languages

Most of the age levels within each scale contain three items, making an approximate total of fifteen across the five scales.

The **DP II** is individually administered, requiring 20 to 40 minutes for administration and scoring. The questions are read from the manual and scored on a separate score form. The "regular" method of administration requires asking all the questions on all the scales. The "short-cut" method is recommended and explained at length in the manual. Essentially it begins with the age level "which makes logical sense based on the age of the child." If the child fails an item at that age level, the examiner proceeds to the next younger level until the child has passed all items in two consecutive age levels (a double basal). Then items are administered until the child fails all items in two consecutive age levels (double ceiling).

The items are written in a question format, which addresses the adult examiner or parent, with limited criteria for yes/no scoring on some items. For example, "Does the child point correctly to at least two colors when asked? The child need not be able to name them." The manual suggests that the questions need not be read exactly as printed, but cautions that they need to be scored according to the exact content of the question. The examiner must be quite familiar with the test content and intent of relevant items in order to be able to paraphrase the questions.

There are directions to determine who will answer the profile questions based on the specific purpose for which the test is used. For example, if a major decision will result from the assessment it is recommended to interview the parent and to test the child directly on those items amenable to direct assessment. For periodic developmental screening the **DP II** can be administered simply as an interview. Regardless of the purpose of assessment, the Socialization and Self-help scales require the respondent to be someone who knows the child well.

Scoring: Items are scored on a pass/fail basis in the scoring booklet, which has an answer sheet for each scale. If the child passes an item, a digit in a "pass" column is circled, indicating how many months credit the child gets for that item. The sum of these circled digits determines the child's developmental age in that skill area. It is somewhat confusing that each "yearly" interval covers from six months prior to six months the given year (e.g., age 5 covers children 4-1/2 to 5-1/2).

The academic scale score can be converted into an "I.Q. Equivalency score" (IQE) by the traditional formula IQE = (Academic Age/Chronological Age) * 100. There are many misinterpretation problems inherent in the use of an IQE, particularly for children who are not represented in the standardization population. The authors recommend that it only be used if such a score is required as a "descriptive label for administrative purposes" or to determine program eligibility.

The manual offers tables of "guidelines" for what would be considered significiant delay, borderline or significantly above normal range. These are based on clinical judgement by individual scale, with no supporting data.

Norms: The norms are rated poor.

The **DP II** was not restandardized when it was revised in 1980. The age norms come from the standardization/tryout completed in 1972. The 1972 standardization sample included data from 3008 children. Only children who met criteria for normality in terms of physical and emotional health were included. Mothers were judged by the raters as to reliability, and data from "unreliable" mothers were discarded. A weekly check was made to maintain representativeness of the sample on the basis of age, sex and race of the child.

The ratio of males to females is fairly even except for the 2 to 2-1/2 year age range where males are seriously overrepresented (63 %) which may have implications for the language-related norms at this age range. Overall, the percent of minorities in the sample is similar to the national population. However, this is not evenly distributed over the age ranges, minorities being more heavily represented in the early years (16-32 %), and dropping to a low of 9 % at age 6. The sample is not evenly distributed on SES nor representative of national SES distributions. It is not geographically representative as the majority of the children were from large cities (Indianapolis, with a small percentage from Seattle), 9% from small cities, and only 2% from rural areas or small towns.

Reliability: The reliability of the **DP II** is rated poor.

Reliability data is reported for very small numbers of mothers and children on the pre-standardization version of the DP. While the percent agreements were high, the small size of the sample, and the fact that there was a limited age range and a limited number of items used, makes the information of very limited value.

More useful, but still limited, information came from the "validity" study comparing 100 mothers' reports to actual testing of children. For the physical scale, 28 of 41 items could be directly observed. The agreement between mother's report and actual testing ranged from 74 to 100% with a mean of 87%. Twenty-one of 48 items on the self-help scale could be directly observed. The agreement between mother's report and actual testing was 77% or greater with a mean of 88%. For the social scale 14 of 45 items could be directly observed. The agreement between mother's report and actual testing ranged from 72 to 100% with a mean of 87%. Twenty-nine of 39 items on the academic scale could be directly observed. The overall agreement between mother's report and actual testing was 86%. For the communication scale, 24 of 44 items could be directly observed. The agreement between mother's report and actual testing averaged 84%.

Validity: Evidence for validity of the **DP II** is rated poor.

Content validity: The content of the **DP II** is virtually identical to the original. The revision was prompted by user requests to modify and clarify the test to meet the provisions for screening under PL 94-142. The test was shortened to eliminate items for ages 10-12. The directions were clarified and items found to be inappropriate, outdated or sexist were eliminated. More specific information about the changes is not provided in the manual.

Individual test items for the original DP were chosen on the basis of established empirical relationships with age reported in the literature, or on other scales of children's intellectual, physical, social and language abilities. A combined item tryout and standardization conducted from 1970 to 1972 provided an update on age placement for the items as well as an elimination or balancing of items with differential age norms by sex or race. (See "Norms" in this review for a description of the issues of representativeness of the standardization sample.) In order to be normative for a particular age an item needed to be passed by 70-80 % of the children in that age range in the standardization sample. The percentage of children passing each item is presented in the manual by sex, race and SES.

A "validity" study was conducted with 100 children ages 3 months to 12 years to determine the reliability of the mother's reports. (The results are described under "Reliability" in this review.) While the level of reliability was high, this study included very low numbers of children by age level and included only an average of 54% of the items on each scale.

Criterion-related validity: Only a few studies are reported which support the validity for only two scales of the original DP. A small study (53 children ages 2 to 11) of the relationship between the physical scale and dental age found signficant correlations between the two for children under 8 years of age. High correlations between academic age and **Stanford Binet** scores supported the IQE, although the sample was small and all children were mentally retarded (and therefore not represented by the age norms). A study of 16 normal children yielded a significant but smaller correlation of .49 between the IQE derived from the academic age and the Binet IQ. This limited evidence of validity was for the original DP and may or may not be true of the **DP II**. No validity evidence is presented for the **DP II**.

Utility: While the concept of the **DP II**, in terms of covering a broad range of skills and utilizing parental reports is a useful one, there are a number of limitations on technical adequacy. The **DP II**, though published in 1980, has normative information from the early 1970s, and which also does not generalize to a broad range of populations. Evidence for validity and reliability is quite limited. There are items for which the scoring criteria are not clear, items that require considerable time to assess directly, or that are impractical to assess directly and items that many children will have had no opportunity to pass (e.g., buying something at the store without help). In addition, some items are very similar from one scale to another (e.g., rhyming is on two scales). A significant distinction is made between questions beginning with "Can the child..."

and those beginning with "Does the child ...", the latter meaning "whether the child *usually* does the task." This leaves room for a level of interpretation that may be difficult for the parent and might lead to inconsistent results.

Use of the "regular" method of screening is questionable since this results in the child being asked to do things that are clearly too easy and things that are clearly too difficult. Considering the attention span of young children and the negative effects of "failing" items, it would seem that the "short-cut" method would be the only method suggested. Most measures that cover a broad age range do establish basal and ceiling levels.

The most positive aspect of the **DP II** is that it provides a structure for the teacher to consider a broad range of children's behavior in terms of strengths and weaknesses. One use for the **DP II** may be for the teacher to interview the parent at the beginning of the year to acquire some familiarity with individual children and later compare their own assessment with the parent's perception of the child.

Availability: Psychological Development Publications, P. O. Box 3198, Aspen, Colorado, 81612.

Instrument: Expressive One-Word Picture Vocabulary Test (EOWPVT, 1979)

Author: Morrison F. Gardner

Purpose: The author's purpose was to provide a quick estimate of a child's expressive verbal intelligence by means of acquired one-word expressive picture vocabulary. The test was designed for children from 2 to 12 years of age.

Description: The **EOWPVT** consists of a book with 110 pages of single pictures. The child names the object on each page. The pictures falling within the 2-8 age range are primarily common objects and some categories (e.g., animals).

The **EOWPVT** is individually administered. It is untimed but the authors report it usually can be administered in 10 to 15 minutes. Because of the range of ages tested, separate starting points have been established for each year of age. Only general directions are given for administration which may limit the consistency of administration procedures.

Scoring: Directions for determining basal and ceiling are clearly presented in the manual. The child's raw score can be converted into the following derived scores: Mental Age, Deviation IQ, Stanine and Percentile Rank. These derived scores are only meaningful in terms of comparing a child's performance to the norms group.

Norms: The norms are rated fair because of the limited geographical representation.

The standardization sample consisted of 1607 children, ages 2-0 through 11-11, residing within the San Francisco Bay area. A statistical weighting procedure was used to ensure that the sample would represent the range and level of ability of children in the United States as much as possible. The sample overrepresented all ethnic groups other than "White" by a small amount, possibly because the authors were concerned about gathering enough data to do bias analyses. No other demographic information is reported.

Reliability: Overall reliability is rated poor, due to limited evidence.

Split-half reliabilities range from .87 to .96, with a median of .94. Test-retest reliability was not examined.

Validity: The validity of the **EOWPVT** is rated fair.

Content validity: The pool of 217 words used for item selection was generated in part from parental reports of children's (ages 18 months to 2 years) word useage. (No information regarding the demographic characteristics of the parents is provided.) Other words were chosen from children's story and textbooks on the basis of face validity to be common within children's homes and not biased by culture, race, sex or bilingual idiosyncracies. Large numbers of children were used for pilot studies to determine the most frequently occuring verbal response to the picture. For some items, more than one response was considered correct, on the basis of these frequency counts. Again, no demographic characteristics are reported for the children participating in these studies, except that they resided in the San Francisco Bay area.

Construct validity: The **Peabody Picture Vocabulary Test** was administered concurrently with the **EOWPVT** to a pilot group of 1,249 children ranging in age from 2 to 11-11. Both item-test correlations and item-**PPVT** correlations were used to determine construct validity. Because the theoretical construct being addressed was language "age," only items were retained which yielded a greater percent passing as chronological age increased were retained.

Criterion-related validity: Either the **Columbia Mental Maturity Scale** or the **Peabody Picture Vocabulary Test** was administered concurrently with the **EOWPVT** to the standardization sample. Correlations with the PPVT ranged from .67 to .78, with a median of .70. Correlations with the **Columbia Mental Maturity Scale** (a measure of general ability) ranged from .29 to .59, with a median of .39. In a separate, pre-kindergarten screening study, correlations of the **EOWPVT** and the subtests of the **Wechsler Preschool and Primary Intelligence Scale (WPPSI)** ranged from .48 to .76, the highest being with the vocabulary subtest.

Utility: The **EOWPVT** is a simple, easily administered, apparently valid test of expressive vocabulary. Under the section on interpreting derived scores, the author states that inferences concerning general ability from this test of expressive vocabulary should be made with caution. However, there are many statements in the manual which imply that the **EOWPVT** can be interpreted as a measure of intelligence.

While there is no evidence to support the use of the **EOWPVT** for these purposes, the manual suggests that this test can provide information about speech defects, possible learning disorders, a bilingual child's fluency in English, auditory processing and auditory-visual association ability. This may be true for a clinican who has other bases for interpreting performance. The manual further suggests that the **EOWPVT** may be used to determine readiness for school or to group children in preschool programs. Based on the evidence provided in the manual these uses are not recommended.

The Spanish translation appears to consist of a direct translation. Translation of the directions is left up to the examiner who should be "fluent" in Spanish. No technical information is available regarding the Spanish version.

Availability: Academic Therapy Publications, 20 Commercial Boulevard, Novato, CA 94947.

Instrument: Human Figures Drawing Test (HFDT, 1986)

Author: Eloy Gonzales

Purpose: The author's purpose is to provide a measure of cognitive maturation in children ages 5 through 10. The test was designed to update previous measures of human figure drawing by responding to the most frequent criticisms in the literature. The results can be used for screening in conjunction with a battery of other tests.

Description: The **HFDT** can be either individually or group administered, requiring 15 to 20 minutes for administation and scoring. The child is asked to make two drawings on a plain sheet of paper, one of themselves and one of someone of the opposite sex. Items are scored on the basis of 38 criteria relating to representation of body parts as described below.

Basic testing techniques for standardized tests are clearly reviewed in the manual. The manual has clear, standardized instructions, with several prompts for the child to "DRAW ALL OF YOURSELF." After the drawings are completed the examiner may probe the child to name any unidentifiable body part.

Scoring: Scoring criteria are clearly presented in the manual (Appendix 2), with examples for most items. For each item a raw score of 1 is awarded if a body part is included in either of the two drawings (self and opposite sex). Some body parts are scored once for presence and again for one or more attributes such as two-dimensional representation, proportionality of trunk or attachment of arms. The total raw score is the sum of all items scored "1" for the 38 items. Raw scores are converted into percentiles and standard scores. The manual provides some information about the interpretation of percentiles and standard scores.

Norms: Normative information is rated good.

The standardization included 2400 public school children as part of a nationally representative sample, stratified by sex, age, geographic region, race and community size. Normative data were collected between September 1982 and January 1985. This sample was consistent with national statistics reported in the 1985 *Statistical Abstract of the United States*. Parental education or occupation was not included as a factor; neither was preschool experience.

Reliability: The reliabilaity of the **HFDT** is rated excellent.

The *internal consistency* (KR-20) reliablity coefficients were .73, .85, .80, .80, .83 and .85 for ages 5 to 10, respectively. Standard errors of measurement (SEM) are reported at 6 to 8 which indicate some lack of confidence in standard scores since two SEMs (95% confidence interval) could make the difference between "poor" and "average" performance.

Test-retest reliability was examined in a separate study of 50 children in grades K, 3 and 5. The **HFDT** was administered twice, with a two week interval between each testing. Reliability coefficients were .87 (kindergarten), .91 (grade 3) and .89 (grade 5).

Reliability of scoring was examined using three examiners who scored the same 30 drawings (10 each of 5-, 8- and 10-year old children from the normative sample). An average inter-scorer correlation of .97 was reported.

Validity: The evidence of validity for the **HFDT** is rated good. *Content validity*: Items were selected on the basis of evidence of developmental progression from the results of other human figures drawing tests. The fact that the percentage of children who pass an item increases with age was considered evidence that items were "developmental." The percentage of children passing each **HFDT** item is presented in the manual for ages 5 to 10 in one-year increments. These percentages (i.e., item difficulties) range from 0 to 100 for five-year-olds and from 10 to 100 for ten-year-olds, indicating an appropriate range of difficulty through the entire age span of the test.

Construct validity: The manual briefly reviews the historical background of the use of human figure drawings in assessment and the controversies surrounding its use as a non-verbal, culture-free measure of intelligence, or more appropriately, cognitive maturation. A brief, rather circular, justification for human figure drawing as a measure of the concept of maturation is presented, based on general normative progressions in children's drawings. Empirical evidence of the construct validity of the **HFDT** was established by the data on age differentiation as well as a study demonstrating predictive relationships between the **HFDT** and academic performance. A study of two groups of children identified as gifted and as mentally handicapped demonstrated that the HDFT scores differentiated among these populations.

Criterion-related validity: Performance on the **HFDT** was compared to concurrent performance on the **Draw-A-Person** test (Harris, 1963), the **Kaufman Assessment Battery for Children** (**KABC**, Kaufman & Kaufman, 1983) and the **WISC-R** (Wechsler, 1974). Thirty students in grades 1, 3 and 5 were given the **HFDT** and the **Draw-A-Person** test with a resultant correlation of .66. Sixty students (ages not reported) were given the **HFDT** and the KABC with a resultant correlation of .52 for the KABC Total score. In a third study, 30 students were given the **HFDT** and the WISC-R with resultant correlations of .53, .31 and .50 for the verbal, performance and full scales, respectively.

Utility: The **HFDT** is a very inexpensive and child-friendly test to administer. No materials are needed other than blank paper and pencils with erasers. The subject matter is familiar and appealing to children and there is no experience of failure even for the youngest children despite the broad age range.

The lack of data on SES or preschool experience in the normative sample leaves some question about the appropriateness of the norms for particular populations of children, although the sample was representative of current population norms on other related factors.

There is strong evidence for the reliability and more limited evidence for the validity of the **HFDT** as a measure of cognitive maturation and ability.

Availability: Pro-Ed, 5341 Industrial Oaks Blvd., Austin, Texas, 78735.

Instrument: Humanics National Child Assessment Form, Ages Three to Six (HNCAF, 1982); Preschool Assessment Handbook (User's Guide, Revised Ed. 1986)

Authors: Derek Whordley, Ph.D., and Rebecca J. Doster

Purpose: The authors' purpose for the **Humanics National Child Assessment Form (HNCAF)** is to provide a checklist of skills and behaviors the child is likely to develop during the ages 3 to 6 years. It is "designed to help the teacher observe the child in different areas of development and to follow changes over the years." The results are to be used for educational planning and not for comparing children or for diagnostic purposes. The form can be used to structure an interview between teacher and parent to discuss children's development, individual characteristics and needs.

The **Humanics National Preschool Assessment Handbook** is designed to inform parents and child development center staff about preschool developmental assessment and to provide information for setting up assessment programs. In addition, it is the user's guide for the **HNCAF** and details specific directions and support materials for using the form to create individualized educational plans and learning activities.

Description: The **HNCAF** includes 90 items which are grouped into five 18-item scales as follows:

Language: receptive and expressive
Cognitive: memory, imagination, thinking, problem-solving
Social-Emotional: cooperation, social awareness, relationship to others, self-concept, expressing and controlling feelings
Motor: gross muscle, fine muscle, visual-motor
Hygiene & Self-Help: recognizing and attending to physical needs, taking responsibility for actions and care of self

Administration: The **HNCAF** is individually administered. Because items may be observed formally or informally over a two-week period, there is no estimation of the time required for administation and scoring. Items are printed on the form and checked as "Occurs Occasionally" if the characteristic or behavior is present but not consistent or firmly mastered; "Occurs Consistently" if a normal part of the child's behavior; or not checked at all. There is space for assessments on four different dates on each form.

A variety of materials is needed to administer the **HNCAF** and a list of materials by item number is provided.

Scoring: The handbook presents a task description/scoring criteria for each item on the **HNCAF**. It is scored as a criterion-referenced test. The Child Development Summary Profile provides a graphic representation of the assessment results. The manual includes general interpretation guidelines.

Norms: The Child Assessment form is not normed.

Reliability: No reliability data was reported for the **HNCAF**.

Validity: The validity of the **HNCAF** is rated good.

Content validity: **The Preschool Assessment Handbook** provides a general discussion of child development between the ages of 3 and 6 years, as well as a brief theoretical framework for the selection of items on **HNCAF**. Behaviors were chosen that "indicate progress in the five developmental areas."

The manual presents a list of indicators that can be considered signs of special problems indicating a child may need further assessment. However, there is no explanation of how these indicators were chosen.

Utility: **The Humanics National Child Assessment Form** is a brief, easily administered checklist that covers a broad range of skills for children ages 3 through 6. It is useful for screening, in that it provides formal documentation of a teacher's observations. It does not have the technical qualities a formal screening test requires (i.e., evidence of validity and reliability).

The handbook describes preschool assessment and the proper use of the **HNCAF** in detail. It provides extensive information about setting up a preschool assessment program including staff training, parent involvement and sample letters to parents. A developmental significance statement is presented with each item, which should provide a better understanding of the item for the examiner and therefore a more accurate assessment.

Availability: Humanics Limited, P.O. Box 7447, Atlanta, Georgia 30309

Instrument: Peabody Picture Vocabulary Test - Revised (PPVT-R, 1981)

Authors: Lloyd M. Dunn, Ph.D., and Leota M. Dunn

Purpose: The authors' primary purpose is to provide a measure of receptive (hearing) vocabulary for Standard American English. The test also yields a quick estimate of verbal ability, although the authors warn that scores are influenced by experiential and cultural factors and should not be interpreted as a direct reflection of "innate" cognitive aptitude.

Description: The **PPVT-R** is a wide-range, norm-referenced power test available in two forms. Forms L and M of the **PPVT-R** each consist of 175 test items, arranged in order of increasing difficulty, preceded by 5 practice items in an easel-bound book. The heaviest concentration of items is for children 3 through 8 years of age. (Because of the staggered starting points, less than 50 items would typically need to be given to any one child.) Each test item consists of a page with four simple, bold line drawings (one correct, three distractors). The child points to the picture that best represents the stimulus word presented orally by the examiner.

The categories of items cover a broad range of topics, including actions, animals, buildings, clothing, foods, things in and about a typical household (e.g., furniture, utensils), human body parts, human workers, plants, shapes, school supplies, tools, toys and vehicles.

The **PPVT-R** is individually administered. It is untimed but it usually can be administered in about 15 minutes. Because of the range of ages that can be tested, separate starting points have been established for each year of age. Easier instructions are given for introducing the test to children under age 8 than for older subjects.. A guide to pronunciation of the stimulus words is provided.

Scoring: The raw score is established by the total of correct responses up to the ceiling item (all responses below the basal are counted as correct). The process of establishing the basal (highest 8 consecutive correct responses) and ceiling (lowest 8 consecutive responses containing 6 errors) is clearly explained in the manual. Tables are given to convert the raw score to a norm-referenced standard score equivalent, percentile rank or stanine. Errors of measurement, definitions and characteristics of the age-referenced derived scores and score range descriptions (e.g., low, moderate, high) are explained in detail in the manual as an aid in interpretation and presenting test results to parents.

Norms: The normative information is rated excellent.

Norms are provided for persons 2 1/2 through 40 years of age. The standardization sample consisted of 4,200 children and youth, very closely matched to the national population (1970 Census) on geographic area, parental occupation, ethnicity and community size. An equivalence study was done which provides information on the correspondence between scores on the **PPVT-R** and the original **PPVT**.

Reliability: The reliability of the PPVT is rated fair.

Split-half *internal consistency* reliability coefficients ranged from .67 to .84 for the 2-8 age groups. *Alternate forms* reliability coefficients ranged from .67 to .83; *test-retest* coefficients from .52 to .78 for the same age range.

Validity: Evidence for the validity of the **PPVT-R** is rated excellent.

Content validity: The selection of test items for the **PPVT-R** was based in part on twenty years of experience and refinements with the original **PPVT**. Information from a number of research studies was used to remove items that were biased culturally, sexually, regionally or racially. The universe of vocabulary from which items were drawn was *Webster's New Collegiate Dictionary* (G & C Merriam, 1957). Items were selected for the **PPVT-R** from a total pool of 700 (144 from the original **PPVT**) on the basis of data from four preliminary tryouts, a calibration study involving a national sample of 5,717 and state-of-the-art item analysis techniques. The item analysis technique allowed test developers to chose items of appropriate difficulty for each age level.

Criterion-related validity: No evidence of criterion-related validity for the **PPVT-R** was available when the manual was published. However, satisfactory evidence is summarized from over 300 validity studies for the original **PPVT**. The correlations were highest with other receptive vocabulary tests (median .86), but the **PPVT** was strongly related to measures of expressive vocabulary (e.g., **EOWPVT**, .70). In addition, **PPVT** has demonstrated moderate correlations with a variety of achievement tests.

Construct validity: Evidence of the **PPVT**'s ability to measure cognitive aptitude is provided in an extensive body of literature. A number of studies have shown that vocabulary is the best single component predictor of intelligence. Construct validitation was also a consequence of the latent-trait item analysis procedure used to scale items.

Utility: The **PPVT-R** is a rigorously developed, psychometrically sound, quickly and easily administered test of receptive vocabulary. It is a convenient, non-threatening and economic way to establish general ability levels for children, with the understanding that it is an aspect of general ability that is heavily influenced by experiential and cultural factors. The test format lends itself to the assessment of language and physically impaired individuals, and as such can be an important part of a diagnostic test battery. The **PPVT-R** is one of the very few tests for young children that has alternative forms, making it useful for situations such as program evaluation involving pre- and post-testing. There is also a Spanish version of the **PPVT-R** available from the publisher. A **PPVT-III** with updated norms is due for release in early 1990.

Availability: American Guidance Service, Circle Pines, Minnesota 55014-1796.

Instrument: Readiness for Kindergarten: A Coloring Book for Parents (1975)

Author: James O. Massey

Purpose: The author's purpose is to provide an activity to help parents determine how ready their child may be for kindergarten.

Description: The test consists of 58 pictured activities labeled, for example, as follows:

12 items	"*Most children entering Kindergarten can ...*" self-help & basic communication skills, count to 10 (ready = 10-11)
12 items	"*Many children entering Kindergarten can ...*" sing, listen, give personal data, colors, button (ready = 8-9)
12 items	"*Half the children entering Kindergarten can ..*" repeat nursery rhymes, hold pencil correctly, use scissors, clap in time to music, understand up, down, etc. (ready = 7-8)
12 items	"*Few children entering Kindergarten can ...*" skip, match rhyming sounds, categorize, copy a square (ready = 7)
10 items	"*Very few children entering Kindergarten can ...*" indicate left/right, print name with upper & lower case, write numbers to 10, read a simple sentence (ready = 5)

Parents mark a box indicating "OK," "?," or "NO." Only the boxes marked OK are used to compare with the ready = number.

The instrument is filled out by parents in their home. If they are concerned about the child's readiness they are referred to school personnel for further discussion. Actually coloring in the book (although the drawings are too detailed for young children) may provide a context for parent-child discussion of kindergarten. This booklet would make a good "transition-to-kindergarten" parent education tool.

Validity: The validity is rated good.

Content validity: The content covers "skills and abilities kindergarten teachers have seen their pupils display within the first month of school." Levels of difficulty were determined by tabulating questionnaires returned by more than 160 experienced kindergarten teachers from schools serving a wide range of socio-economic areas.

Utility: The book provides a good format for readiness awareness and discussions with parents. The booklet also provides some interpretation and suggestions for activities parents can do with their children to prepare them for kindergarten. While the skill content of the book is not necessarily dated, the recommendation at the end of the book, that it is better to keep children out of school if they do not appear to be ready, is based on an outdated "maturational" concept of readiness. Teachers should be aware of that and use it as a discussion point for parents.

Availability: Consulting Psychologists Press, Palo Alto, CA

Instrument: Receptive One-Word Picture Vocabulary Test (ROWPVT, 1985)

Author: Morrison F. Gardner

Purpose: The author's purpose is to provide an assessment of a child's "one-word hearing vocabulary." This test was developed as a companion test to the .**Expressive One-Word Picture Vocabulary Test (EOWPVT)** to provide comparable normative information on receptive vocabulary. The author does not provide a rationale for the use of this test in the framework of measuring language ability.

Description: The test consists of 100 test plates representing vocabulary words ordered in respect to difficulty. The child indicates the picture that represents the word presented orally by the examiner. The **ROWPVT** is individually administered. It is untimed but the authors report it usually can be administered in less than 15 minutes. Because of the range of ages tested, separate starting points have been established for each year of age. Only general directions are given for administration which may limit the consistency of administration procedures. There is a pronunciation guide in the manual for all the words on the test.

Scoring: Raw scores can be converted into four types of derived scores: language age, language standard score, stanine and percentile rank. The fact that these scores indicate a child's standing relative to the normative sample is of limited applicability because of the sample characteristics described below. This same factor limits the comparison of percentile ranks on the **ROWPVT** with percentile ranks on other tests.

Norms: Normative information is rated fair.

This test was not really normed. The standardization sample consisted of 1128 children, ages 2-0 through 11-11, residing within the San Francisco Bay area. No information is given about the demographics of the standardization sample. This and the fact that it was **not** a representative sample in terms of the national population limits the applicability of the "norms" and of the derived scores as noted above.

The authors justify the lack of representativeness of the sample on the basis that the **ROWPVT** was scaled using the **EOWPVT** which had a more representative sample. A concurrent administration of the **EOWPVT** was used to equate scores between the two measures. The author uses the term "equivalent" to describe the norms of the **ROWPVT** and the **EOWPVT**. "Comparable" is the proper term. Scores that are equivalent measure the *same* trait.

Reliability: Reliability of the **ROWPVT** is rated poor because of limited evidence

Split-half reliabilities range from .81 to .93, with a median of .90. Test-retest reliability was not examined.

Validity: Evidence of the validity of the **ROWPVT** is rated fair.

Content validity: Six hundred pictures were selected to represent a common core of English words familiar to children in the home or school environment. These were reviewed for face validity in terms of age appropriateness by teachers from preschool and grades K-6, as well as language and speech pathologists. An effort was made to eliminate pictures that might be regionally, ethnically, culturally or sex biased. The resultant set of 150 was reduced to a final set of 100 on the basis of a pilot study (415 children, ages 2-0 to 11-11) and subsequent item analysis.

Criterion-related validity: The vocabulary subtest of the WPPSI or WISC-R was given to 935 of the children in the standardization sample with resultant validity coefficients decreasing from .70 to .42 for the 4-0 to 5-6 age groups on the WWPSI and varying from .23 to .41 among the six age groups (6-0 to 11-0) on the WISC-R. Construct validity was more strongly supported through the relationship between the **ROWPVT** and the **EOWPVT** (r=.89). It is interesting that the author did not chose the PPVT-R as a criterion since the tests measure virtually the same thing.

Comment: The **ROWPVT** is a simple, easily administered test of receptive vocabulary. Other than its relationship with the **EOWPVT**, it does not have any attributes that give it an advantage over the much more rigorously developed, and equally easily administered, PPVT-R. The use of "norms" is deceptive since a casual user may not understand the serious limitation of the way the **ROWPVT** was "standardized." The Spanish translation appears to consist of a direct translation. Pictures and words that could not be translated into Spanish were eliminated in the development process for the English version. Translation of the directions is left up to the examiner who should be "fluent" in Spanish. The Spanish version has not been standardized.

Availability: Academic Therapy Publications, 20 Commercial Boulevard, Novato, CA 94947.

APPENDIX H

SUMMARY TABLE OF OTHER EARLY CHILDHOOD INSTRUMENTS

Content and Key to Instrument Descriptors in Review Summary Tables

INSTRUMENT: *Instrument name, acronym, author(s), publication date and publisher. Indices of instruments by title and publishers' addresses are included after Appendix J.*

FOCUS: *Scope of content covered by the instrument.*

Broad: Includes three or more of the following categories of abilities:
Language, Speech, Cognition, Perception, Personal/Social, Perceptual-motor, Fine, Gross Motor Coordination
Academics: Includes many, but primarily academic skills
Specific Areas: Language, Literacy, Mathematics, Reading, Relational Concepts (see "Content" for specific skills in each area)

AGE/GRADE: *Age or grade range covered by the instrument.*

ADM. TIME: *Time in minutes required for administration and initial scoring.*

FORMAT: *Description of test in terms of type of response required, format and materials, categories are not mutually exclusive*

Format: Group or Individual Administration
Multiple choice
Paper & Pencil (child marks or writes the answer)
Stimulus cards/easel
Manipulatives (e.g., blocks, sorting chips)
Response Mode : Teacher rating
Parent response
Observation of Child
Oral (verbal)
Pointing (implies multiple choice)
Performance (*fine/visual-motor*: copy, build, write, etc)
Motor (*gross motor*: hop, skip, jump, catch, etc.)

SCORES: *Types of scores available. No endorsement of the use of specific types of scores is implied* here.

Norm-referenced: Percentile, Percentile Rank
Age Equivalent / Grade Equivalent (Gr.Eq.)
Standard Score
Normal Curve Equivalent (NCE)
Developmental "Age", "Language Age", etc.
Quotient (Developmental, Language, etc.)
Criterion-referenced: Mastery levels
Raw score

CONTENT: *When the content covers a number of areas, the category name is used. When the content is more limited within a category, the specific areas are named.*

Basic facts: colors (primary), letters, numbers. shapes
Language: expressive, receptive vocabulary, fluency, syntax
Literacy: print functions & conventions, reading symbols
Relational Concepts: direction, position, size, quantity, order, time, categorization
Listening & Sequencing: follows directions, remembers story sequences, main ideas
Cognitive: problem solving, opposite analogies, memory, imitation
Perception: auditory, visual discrimination
Mathematics: count rote, with 1/1 correspondence, number skills
Motor: fine motor (holding a pencil correctly, buttoning, etc)
gross motor (hops, skips, throws)
visual-motor (copies shapes, builds blocks)
Self: knowledge of body parts (point or name)
social/emotional (peer & teacher interactions, attention span, etc.)
self help (buttoning, toilet, etc)
information (name, age, address, phone, birthdate)

NORMS: *Ratings on norming studies (value judgement implied)*

None: no normative information is given
Poor: some information but limited applicability
Fair: some standards of comparison (e.g., means of research sample)
Good: norms based on good sized, representative sample,
or lots of relevant information regarding appropriate populations for use
Excellent: norms based on a representative, national sample and relevant
information about applying norms or norm-referenced scores.

RELIABILITY: *Reliability ratings (value judgement implied)*

None: no reliability information is provided
Poor: all reliability coefficients (r) below .70
or an important type of reliability was not examined
Fair: at least one reported r is greater than .70; or r was
greater than .80 but evidence was limited in applicability
Good: total r is greater than .80; most subtests have r greater than .75
Excellent: several kinds of reliability reported; total r is greater
than .90; most subtest scores greater than .80

VALIDITY: *Validity ratings (value judgement implied)*

None: no validity information is provided
Poor: information is of very limited applicability
Fair: most important aspects of were addressed but evidence was
moderate or weak; or was strong but limited in applicability
Good: consistent evidenct of validity, or strong but limited evidence
of the type of validity most appropriate for the intended test use
Excellent: strong evidence and a base of research on the instrument

Summary Table of Instrument Characteristics: Other Early Childhood Measures

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Battelle Developmental Inventory (BDI) 1984 DLM Teaching Resources	Developm. Inventory	Ages 0 - 8	90 - 120 (ages 3 - 5)	Individually Adm Sprial bound Oral Motor	Self Motor Cognitive Language	Standard Percentile	Fair	Excellent	Good	Instructions for IEP development Specific adaptations for handicapped children
Diagnostic Inventory of Early Development (IED) Brigance, 1978 Curriculum Associates, Inc	Developm. Inventory	Ages 0 - 7	untimed	Individually Adm Oral Performance	Reading readiness Language Mathematics	Criterion Referenced No summary	None	None	Fair	"Norms" for items from published texts and curriculum materials
Diagnostic Inventory of Basic Skills (IBS Brigance, 1977 Curriculum Associates , Inc	Developm. Inventory	Grades K - 6	untimed	Individually Adm Oral Performance	Self Motor Cognitive Lang & Math	Criterion Referenced No summary	None	None	Fair	"Norms" for items from published develomental norms
Developmental Profile II (DPII) Alpern, Boll & Shearer, 1980 Psychological Development Publications	Developm. Inventory	Ages 0 - 9	20 - 40	Individually Adm Motor Oral Performance	Self Motor Basic Facts Language	Devel. Age by area IQ Equiv.	Poor	Poor	Poor	
Expressive One Word Picture Vocabulary Test (EOWPVT) Gardner, 1979 Academic Therapy Publications	Language	Ages 2 - 12	10 - 15	Individually Adm Stimulus cares Oral	Picture vocabulary expressive	Percentile Mental age Deviatn IQ Stanine	Fair Limited	Poor Limited	Fair	
Human Figures Drawing Test (HFDT) Gonzales, 1986 PRO-ED	Cognitive Maturation	Ages 5 - 10	15 - 20	Individually Adm Drawing	Draw self & person of opposite sex	Percentile Standard	Good	Excellent	Good	No validity as a readiness test
Humanics National Child Assessment Form, Ages 3 -6 Whordley & Doster, 1982 (HNCAF) PRO-ED	Develop. Inventory	Ages 3 - 6	untimed	Individually Adm Observational Checklist	Language Cognitive Self Motor	Criterion Referenced Summary Profile	None	None	Good	Preschool Assessment Handbook accompanies;

Summary Table of Instrument Characteristics: Other Early Childhood Measures cont.

INSTRUMENT	DESCRIPTION						TECHNICAL QUALITY			
	Focus	*Ages/ Grades*	*Adm. Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
Peabody Picture Vocabulary Test, Revised (PPVT-R) Dunn & Dunn, 1981 American Guidance Service	Language	Ages 2 to adult	15	Individually Adm Stimulus easel Oral	Picture vocabulary receptive	Percentile Standard Stanine	Excellent	Fair	Excellent	The standard for this type of test. Used in a very large number of research studies
Readiness for Kindergarten: A coloring Book for Parents Massey 1975 Consulting Psychologists Press	Language	Grade PreK	untimed	Parent Observation Checklist	Picture vocabulary receptive	Percentile Lang. age Standard Stanine	None	None	Good	Somewhat outdated concept of readiness but may be used to communicate with parents
Receptive One Word Picture Vocabulary Test (ROWPVT) Gardner, 1985 Academic Therapy Publications	Language	Ages 2 - 12	15	Individually Adm Stimulus cards Oral	Picture vocabulary receptive	Percentile Lang. age Standard Stanine	Fair	Poor	Fair	

Summary Table of Instrument Characteristics: Achievement Batteries

INSTRUMENT	DESCRIPTION					TECHNICAL QUALITY			
	Ages/ Grades	*Adm Time*	*Format*	*Content*	*Scores*	*Norms*	*Reliability*	*Validity*	*Comment*
California Achievement Tests (CAT E/F) CTB/McGraw-Hill, 1985	Grades K - 12	150	Group Adm Multiple Choice Paper & Pencil	Visual & Sound Recognition Vocab. Oral Comprehension Language Expression Math Concpets & Applicaitons	Scale Scores Percentiles NCE, Gr.Eq. Stanines	Excellent	Fair	Fair	Curriculum referenced also Classroom management guide includes instructional activites
Gates-MacGinitie Reading Tests MacGinitie, 1978 The Riverside Publishing Company	Grades K - 12	55	Group Adm Multiple Choice Paper & Pencil	Vocabulary Comprehenison	Descriptive Low/High/Avg (lowest level	Fair Dated	Good	Fair	
Iowa Tests of Basic Skills (ITBS) Hieronymus, Hoover & Lindquist, 1986 The Riverside Publishing Company	Grades K - 9	160	Group Adm Multiple Choice Paper & Pencil	Listening, Word recognition Vocabulary, Word Analysis Reading Comprehension Language & Math Skills	Grade Eq. Scale scores	Excellent	Fair	Fair	Seven separate sets of norms including large city, Catholic schools and high/low SES
Metropolitan Achievement Tests (MAT6) The Psychological Corporation	Grades K - 12	95	Group Adm Multiple Choice Paper & Pencil	Reading Math, Language, Vocabualry, Word Recognition Reading Comprehension	Gr. Eq., NCE Percentiles Scale Score	Good	Fair	Fair	Survey & Diagnostic forms Aslo provides criterion-referenced scores
Peabody Individual Achievement Test Dunn & Markwardt, 1970 (PIAT) American Guidance Service	Grades K - 12	30 - 40	Individually Adm Easel kits	Math, Reading Recognition Comprehension, Spelling General Informaiton	Age & Gr. Eq. Percentiles Standard	Dated Good	Good	Limited Poor	Easel format has stimulus pictures on one side and instructions on the other
Stanford Early School Achievement Test; Madden, Gardner & Collins, 1983 The Psychological Corporation (SESAT)	Grades K & 1	130	Group Adm Multiple Choice Paper & Pencil	Sounds & Letters Word Reading Listening toWords & Stories Math, Environment	Stanines Grade Eq. Percentiles Standard	Good	Fair	Fair	Standardized at midyear only Attractive format
SRA Achievement Series Naslund, Thorpe & Lefever, 1978 Science Research Associates	Grades K - 12	120	Group Adm Mulitple Choice Paper & Pencil	Vis & Aud Discrimination, Letters & Sounds, Listening Math Concepts	Gr.Eq. NCE Percentiles Stanines	Good	Good	Good	Includes some criterion-referenced information
Wide Range Achievement Test Jastak & Wilinson, 1984 (WRAT-R) Jastak Assessment Systems	Ages 5 - 12 12 - 74	15 - 30	Individually Adm Paper & Pencil Some Performance	Reading Spelling Arithmetic	Grade Eq. Percentiles Standard	Fair	Unclear	Fair	

APPENDIX I

CODE OF FAIR TESTING

CODE OF FAIR TESTING PRACTICES IN EDUCATION

Prepared by the Joint Committee on Testing Practices

The Code of Fair Testing Practices in Education states the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of testing. Although the Code has relevance to many types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs. The Code is not intended to cover tests made by individual teachers for use in their own classrooms.

The Code addresses the roles of test developers and test users separately. Test users are people who select tests, commission test development services, or make decisions on the basis of test scores. Test developers are people who actually construct tests as well as those who set policies for particular testing programs. The roles may, of course, overlap as when a state education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

* * *

The Code has been developed by the Joint Committee on Testing Practices, a cooperative effort of several professional organizations, that has as its aim the advancement, in the public interest, of the quality of testing practices. The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. In addition to these three groups, the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association are now also sponsors of the Joint Committee.

This is not copyrighted material. Reproduction and dissemination are encouraged. Please cite this document as follows:

Code of Fair Testing Practices in Education. (1988) Washington, D.C.: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, American Psychological Association, 1200 17th Street, NW, Washington, D.C. 20036.)

The Code presents standards for educational test developers and users in four areas:

A. Developing/Selecting Tests
B. Interpreting Scores
C. Striving for Fairness
D. Informing Test Takers

Organizations, institutions, and individual professionals who endorse the Code commit themselves to safeguarding the rights of test takers by following the principles listed. The Code is intended to be consistent with the relevant parts of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985). However, the Code differs from the Standards in both audience and purpose. The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests. The Code is not meant to add new principles over and above those in the Standards or to change the meaning of the Standards. The goal is rather to represent the spirit of a selected portion of the Standards in a way that is meaningful to test takers and/or their parents or guardians. It is the hope of the Joint Committee that the Code will also be judged to be consistent with existing codes of conduct and standards of other professional groups who use educational tests.

A Developing/Selecting Appropriate Tests*

Test developers should provide the information that test users need to select appropriate tests.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations.

Test Developers Should:	**Test Users Should:**
1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.	1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.
2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended purposes.	2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).	3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Describe the process of test development. Explain how the content and skills to be tested were selected.	4. Become familiar with how and when the test was developed and tried out.
5. Provide evidence that the test meets its intended purpose(s).	5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.	6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.	7. Ascertain whether the test content and norms group(s) or comparison group(s) are appropriate for the intended test takers.
8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.	8. Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

---

*Many of the statements in the Code refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the test development process should be designed to help ensure that the completed tests will be in compliance with the Code.

B Interpreting Scores

Test developers should help users interpret scores correctly.

Test Developers Should:

9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.

10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.

11. Warn users to avoid specific, reasonably anticipated misuses of test scores.

12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.

13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).

Test users should interpret scores correctly.

Test Users Should:

9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.

10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.

11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.

12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.

13. Obtain evidence to help show that the test is meeting its intended purpose(s).

C Striving for Fairness

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Developers Should:

14. Review and revise test questions and related materials to avoid potentially insensitive content or language.

15. Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.

16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Users Should:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.

15. Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test.

16. When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

D Informing Test Takers

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

Test Developers or Test Users Should:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.

18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers.

Under some circumstances, test developers have direct control of tests and test scores. Under other circumstances, test users have such control. Whichever group has direct control of tests and test scores should take the steps described below.

Test Developers or Test Users Should:

19. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.

20. Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.

21. Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■

Note: The membership of the Working Group that developed the Code of Fair Testing Practices in Education and of the Joint Committee on Testing Practices that guided the Working Group was as follows:

Theodore P. Bartell
John R. Bergan
Esther E. Diamond
Richard P. Duran
Lorraine D. Eyde
Raymond D. Fowler
John J. Fremer
  (Co-chair, JCTP and Chair, Code Working Group)
Edmund W. Gordon
Jo-Ida C. Hansen
James B. Lingwall
George F. Madaus
  (Co-chair, JCTP)
Kevin L. Moreland
Jo-Ellen V. Perez
Robert J. Solomon
John T. Stewart
Carol Kehr Tittle
  (Co-chair, JCTP)
Nicholas A. Vacc
Michael J. Zieky
Debra Boltas and Wayne Camara of the American Psychological Association served as staff liaisons

Additional copies of the Code may be obtained from the National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, D.C. 20036. Single copies are free.

[illegible] • [illegible] • [illegible] • Printed in U.S.A.

APPENDIX J

REFERENCE WORKS FOR EARLY CHILDHOOD ASSESSMENT

REFERENCE WORKS FOR EARLY CHILDHOOD ASSESSMENT

Bate, Margaret, Smith, M., and James, J. (1981). *Review of tests and assessments in early childhood*. Atlantic Highlands, NJ: Humanities Press, Inc.

Beaty, Janice J. (1986). *Observing development of the young child*. Columbus, OH: Charles E. Merrill Publishing Company.

Cross, A.W. (1985). Health screening in the schools. *The Journal of Pediatrics*, 107:487-494, 653-660.

Goodwin, W. L., and Driscoll, Laura A. (1980). *Handbook for measurement and evaluation in early childhood education*. San Francisco, CA: Jossey-Bass Publishers.

ETS Test Collection (1987). *Criterion-referenced measures, preschool - grade 3*. Princeton, NJ: Educational Testing Service.

Frankenburg, W. K., Emde, R. N., and Sullivan, J. W. (Eds.) (1985). *Early identification of children at risk*. New York, NY: Plenum Press.

Keyer, Daniel J., and Sweetland, Richard C. (Eds.) (1984-1987). *Test critiques, Volumes I - VI*. Kansas City, MO: Test Corporation of America, a Subsidiary of Westport Publishers, Inc.

Meisels, S.J. (1985). Developmental screening in early childhood: A guide (rev. ed.). Washington, DC: National Association for the Education of Young Children.

Meisels, S. J., and the Expert Team on Screening and Assessment, NCCIP (1988). Guidelines for the identification and assessment of young disabled and developmentally vulnerable children and their families. National Center for Clinical Infant Programs, National Early Childhood Technical Assistance System.

Minnesota Department of Education (1985). *Instruments and procedures for assessing young children*.

Salvia, J. & Ysseldyke, J.E. (1988). *Assessment in special and remedial education, Fourth Edition*. Boston, MA: Houghton Mifflin Company.

Sattler, J.M. (1988). *Assessment of children, Third edition*. San Diego, CA: Jerome M. Sattler, Publisher.

Schakel, Jacqueline, and Duthie, Jill (1986). *Assessment manual for preschool special education*, Preschool Resources for Alaskan Special Education.

Index of Instruments by Category

APPENDIX A: SCREENING

Page		
1	BSSI-S	Basic School Skills Inventory - Screening
*	BDI-S	Battelle Developmental Inventory - Screening
3	BBCS-S	Bracken Basic Concept Scale, Screening Forms
6		Brigance K & 1 Screen
9		Brigance Preschool Screen
*		The Communication Screen
*	DDST	Denver Developmental Screening Test
11	DASI II	Developmental Activities Screening Inventory-II
13	DIAL-R	Developmental Indicators for the Assessment of Learning- Revised
17	EISP	Early Identification Screening Program
19	ESI	Early Screening Inventory
22	FKSB	Florida Kindergarten Screening Battery
*		Fluharty Preschool Speech and Language Screening Test
24	KLST	Kindergarten Language Screening Test
*	MST	McCarthy Screening Test
26	MAP	Miller Assessment for Preschoolers
*	MSEL	Mullen Scales of Early Learning
30	PEER	Pediatric Early Examination of Readiness
*	SCREEN	Screening for Related Early Educational Needs
*		SEARCH: A Scanning Instrument for the Identification of Potential Learning Disability

APPENDIX D: MASTERY OF READINESS (EARLY ACHIEVEMENT) CONCEPTS

Page		
1		Analysis of Readiness Skills
4	BSSI-D	Basic School Skills Inventory Diagnostic
7	Boehm-R	Boehm Test of Basic Concepts-Revised
11	Boehm-PV	Boehm Test of Basic Concepts-Preschool Version
15	BBCS-D	Bracken Basic Concept Scale, Diagnostic
*		CIRCUS
19	CSAB	Cognitive Skills Assessment Battery
*		Gesell Preschool Test
*	SRST	Gesell School Readiness Test, School Readiness Screening Test
21		The Lollipop Test
*	MRT	Metropolitan Readiness Tests, 1986 Edition
*	PI	Preschool Inventory
24	SRS	School Readiness Survey
*	TOBE 2	Tests of Basic Experiences 2
26	TELD	Test of Early Language Development
28	TEMA	Test of Early Mathematics Ability
30	TERA	Test of Early Reading Ability
*	TEWL	Test of Early Written Language
32	TOLD-2	Test of Language Development, Primary

* No full review, brief review in summary tables, Appendix C (Screening) or F (Readiness)

APPENDIX G: OTHER EARLY CHILDHOOD MEASURES

Developmental Inventories

Page		
*	BDI	Battelle Developmental Inventory
1	DP II	Developmental Profile II
*	IED	Diagnostic Inventory of Early Development (Brigance)
*	IBS	Diagnostic Inventory of Basic Skills (Brigance)
9	HNCAF	Humanics National Child Assessment Forms

Cognitive Maturity

5	EOWPVT	Expressive One-Word Picture Vocabulary Test
7	HFDT	Human Figures Drawing Test
11	PPVT-R	Peabody Picture Vocabulary Test-Revised
14	ROWPVT	Receptive One-Word Picture Vocabulary Test

Miscellaneous

13		Readiness for Kindergarten: A Coloring Book for Parents

Achievement Batteries

*	CAT E/F	California Achievement Tests, Forms E and F
*		Gates-MacGinitie Reading Tests
*	ITBS	Iowa Tests of Basic Skills
*	MAT6	Metropolitan Achievement Tests
*	PIAT	Peabody Individual Achievement Test
*	SRA	SRA Achievement Series
*	SESAT	Stanford Early School Achievement Test
*	WRAT	Wide Range Achievement Test

* No full review, brief review in summary table, Appendix H

LIST OF PUBLISHERS WITH ADDRESSES

Academic Therapy Publications, 20 Commercial Boulevard, Novato, CA 94947-6191

American Guidance Service, Publishers' Building, Circle Pines, MN 55014-1796

Behavior Science Systems, P.O. Box 1108, Minneapolis, MN 55440

Childcraft Education Corporation, 20 Kilmer Road, P.O. Box 3081, Edison, NJ 08818-3081

Communication Skill Builders, Inc., 3130 N Dodge Blvd., P.O.Box 42050, Tucson, AZ 85733

Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, CA 94306

CTB/McGraw-Hill, 2500 Garden Road, Monterey, CA 93940

Curriculum Associates, Inc., 5 Esquire Road, North Billerica, MA 01862-2589

DLM Teaching Resources, One DLM Park, P.O. Box 4000, Allen, TX 75002

Educational Testing Service, Rosedale Road, Princeton, NJ 08541

Educators Publishing Service, Inc., 75 Moulton Street, Cambridge, MA 02238-9101

Foundation for Knowledge in Development, 11715 East 51st Avenue, Denver, CO 80239

Humanics, Limited, 1182 West Peachtree Street, Suite 201, Atlanta, GA 30309

Jastak Associates, Inc., 1526 Gilpin Avenue, Wilmington, DE 19806

LADOCA Publishing Foundation, 5100 Lincoln Street, Denver, CO 80216

Modern Curriculum Press, 13900 Prospect Road, Cleveland, OH 44136

PRO-ED, 5341 Industrial Oaks Blvd., Austin, TX 78735

Psychological Assessment Resources, Inc., P.O. Box 998, Odessa, FL 33556

The Psychological Corporation, 555 Academic Court, San Antonio, TX 78204

Psychological Development Publications, P.O. Box 3198, Aspen, CO 81612

Programs for Education, Inc., Department W-16, 82 Park Avenue, Flemington, NJ 08822

The Riverside Publishing Co., 8420 Bryn Mawr Avenue, Chicago, IL 60631

Science Research Associates, Inc., 155 North Wacker Drive, Chicago, IL 60606

Teachers College Press, Teachers College, Columbia University, 1234 Amsterdam Avenue, New York, New York, 10027

T.O.T.A.L. Child, Inc., 244 Deerfield Road, Cranston, RI 02920

Walker Educational Book Corporation, 720 Fifth Avenue, New York, NY 10019

THE TEST CENTER

The Test Center at the Northwest Regional Educational Laboratory is a library of tests and testing resources. Materials are loaned to educators in Alaska, Hawaii, Idaho, Montana, Oregon, Washington and the Pacific Island. Many of the early childhood instruments in this guide are available for a three week loan by contacting:

The Test Center
Northwest Regional Educational Laboratory
101 SW Main Street, Suite 500
Portland, OR 97204
503/275-9500 or 800/547-6339